약학박사 학위논문

직접주입-질량분석법 기반의 대사체학 및 기계학습을 통한 원산지 판별분석법

Direct Infusion-Mass Spectrometry based Metabolomics and Machine Learning for the Discrimination of Geographical Origin

2018년 02월

서울대학교 대학원

약학과 약품분석학전공

임 동 규

직접주입-질량분석법 기반의 대사체학 및 기계학습을 통한 원산지 판별분석법

Direct Infusion-Mass Spectrometry based Metabolomics and Machine Learning for the Discrimination of Geographical Origin

지도교수 권 성 원

이 논문을 약학박사 학위논문으로 제출함

2018년 02월

서울대학교 대학원

약학과 약품분석학전공

임 동 규

임동규의 박사학위논문을 인준함

2018 년 02 월

위 원 장 박 정 일 (인)

부 위 원 장 이 용 문 (인)

위 원 나 동 희 (인)

위 원 임 요 한 (인)

위 원 권 성 원 (인)

Abstract

본 연구는 Direct Infusion-Mass Spectrometry (DI-MS)와 기계 학습을 대사체학 연구에 도입하여 컬럼 크로마토그래피 및 다변량 통계분석을 주로 이용하는 기존의 방식보다 분석 시간, 판별의 정확도 측면에서 대폭 개선된 판별 분석 전략을 제시한다. 제시한 방법의 포괄적인 적용 가능성을 증명하기 위해 원산지가 다른 2 종의 쌀(한국산 쌀, 중국산 쌀 및 임의의 비율로 이를 혼합한 쌀, 총 430 점) 및 품종이 다른 4 종의 인삼(*P. ginseng, P. notoginseng, P. quinquefolius, P. Vietnamensis,* 총 40 점)을 대상으로 각각 서로 다른 분석 장비(쌀: Agilent triple quadrupole 6460 system, 인삼: WatersMirco triple quadrupole mass spectrometer)를 이용하여 판별 분석을 진행하였다. Liquid chromatography-mass spectrometry(LC-MS) 결과를 통해 확인된 판별 대사체들은 쌀의 lysophospholipid 와 인삼의 ginsenoside 로 확인되었으며, 확인된 판별 대사체를 대상으로 시료 당 30 초 내외의 분석시간 및 우수한 감도를 제공하는 DI-MS 기반의 분석법을 이용해 동시 분석하였다. 이후 해당 확보한 데이터를 토대로 판별 모델을 제작하기 위해 다변량 통계분석인 principal component analysis, partial least squares discriminant analysis 와 기계 학습 방법인 random forest, support vector machine, C5.0 decision tree, neural network 및 k-nearest neighbor classifiers 를 적용하였으며 판별 모델을 서로 비교, 평가하였다. 판별 분석 결과, 특히 기계 학습을 이용하였을 때 2 종의 쌀 및 4 종의 인삼이 훨씬 정확하게 판별되었으며 특히 기계학습 방법 중 하나인 random forest 는 다변량 통계분석으로 판별해내지 못하는 5% 내외로 혼입된 혼합 쌀 시료도 판별해 낼 만큼 매우 정밀한 판별 결과를 보여주었다. 결론적으로 대사체학 연구, 특히 판별 연구에 있어 DI-MS 와 기계 학습을 도입한 판별 분석 전략은 대량의 시료, 신속한 속도, 뛰어난 정확도를 요구하는 연구를 위한 매우 적합한 방법이다.

주요어: 판별 분석, Direct infusion-mass spectrometry, 기계 학습

학번: 2012-21614

Contents

List of figures

List of tables

1. Introduction

특정 시료에 존재하는 대사체(metabolite)를 분석하는 대표적인 방법으로써 대사체학(metabolomics)은 1)질병의 진단, 2)의약품의 성분 검출, 3)영양 성분의 확인, 4)식물생물학 등 수많은 분야에서 다양하게 응용 및 적용되고 있다 [1]. 특히 대사체학 기반의 분석을 통해 수천개에 이르는 대사체를 확인하고 각 대사체별로 상대적인 함량을 비교 가능하기에 이를 응용한 판별 분석 연구에도 활발하게 활용되고 있다 [2, 3]. 이러한 판별 분석을 통해 특정 대사체의 변화와 외부 자극 요인(외부 환경의 변화, 의약품의 섭취 등)의 상관관계를 알아낼 수 있으며 나아가 대사체간의 metabolic network 및 metabolic pathway 를 알아내는 것 또한 가능하다 [4]. 이와 같은 학술적인 분야 뿐 아니라 대사체학 기반의 판별 분석은 국제 식품 정책의 다변화에 따른 엄격한 품질관리의 필요성으로 인해 식품, 약용작물의 원산지 판별, 혼입 유무 판별, 오염 판별 등의 실용적인 부분에서도 매우 활발하게 이용되고 있다 [5].

판별 분석을 위한 포괄적인 대사체 분석을 위해서는 일반적으로 특별히 분석 대상을 정하지 않고 모든 대사체를 분석하는 untargeted metabolomics 를 이용한 접근법을 사용한다 [6, 7]. 대사체를 분석, 규명 및 정량적으로 측정 가능한 기기는 기본적으로 모두 대사체학에 응용 가능하지만 대사체학에서 가장 일반적으로 사용되는 분석 기법은 핵자기공명법(nuclear magnetic resonance spectroscopy, NMR)와 질량분석법(mass spectrometry, MS)이다. NMR 은 대표적인 비파괴 분석방법으로써 매우 높은 재현성(reproducibility) 및 정량성을 보여주며 매우 넓은 분석 범위를 가지고 있다. 무엇보다 NMR 은 성분의 구조를 정확하게 파악 가능하므로 stable isotope labels 등을 이용하여 특정 대사체의 변화 과정을 추적하기에 매우 적합하며, 따라서 약물의 체내 동태 변화 등의 대사체 추적, 스크리닝 연구에 활발하게 사용된다 [8].

하지만 NMR 은 매우 낮은 민감도(sensitivity)를 가지고 있기에 미량 대사체 분석에 일반적으로 적합하지 않다. 또한 NMR 단독으로는 대사체 간의 분리가 사실상 불가능 하기에 원하는 특정 대사체만을 선택적으로 검출, 분석하는 것이 어려우며 이는 특정 대사체의 함량을 통한 판별 분석을 위해서는 적절하지 않은 특징이다 [9]. MS 는 매우 우수한 정확도(accuracy), 민감도(sensitivity), 넓은 분석 범위 (dynamic range)를 가지고 있으며 특히 분자량으로 대표되는 대사체 구명의 용이함으로 인해 오늘날 대사체학에서 가장 많이 사용되는 분석 기법이다 [10]. MS 는 일반적으로 두가지 대표적인 성분 분리 방법인 기체 크로마토그래피(gas chromatography, GC), 및 액체 크로마토그래피(liquid chromatography, LC)와 함께 사용된다. GC-MS 는 아미노산, 당, 당알코올과 같은 1 차 대사체(primary metabolite)를 분석하는데 매우 적합한 분리 방법으로 매우 우수한 머무름 시간(retention time) 및 스팩트럼(mass spectrum)의 재현성을 가지고 있다 [11]. 무엇보다 GC-MS 는 뛰어난 재현성으로 인해 National Institute of Standards and Technology (NIST) 와 같은 라이브러리가 높은 수준으로 구현되어 있으며 이를 통해 미지 대사체 구명에 매우 유리하다. 하지만 GC-MS 의 경우 휘발성 물질들만 분석이 가능하며 경우에 따라 시료 준비 과정에서 복잡한 유도체화 과정을 거쳐야만 분석이 가능하다는 단점 또한 가지고 있다. 반면 LC-MS 는 GC-MS 와 달리 분석 범위의 한계가 사실상 없으며 primary metabolite 와 함께 지질, 알칼로이드, 플라보노이드와 같은 2 차 대사체(secondary metabolite) 또한 검출이 가능하다 [7]. 하지만 GC-MS 에 비해 상대적으로 재현성이 떨어지며 이러한 이유로 대사체 구명에 어려움이 따른다. 그럼에도 불구하고 LC-MS 가 어떠한 분석 방법보다도 다양한 종류의 대사체를 검출해내는 것이 가능하기 때문에 특히 고유한 2 차 대사체를 가지는 식물체의 판별 분석에 있어서 LC-MS 는 가장 이상적인 플랫폼이라고 할 수 있다 [7, 12].

위와 같이 GC 또는 LC 와 같은 성분 분리 방법을 도입할 경우 매트릭스 효과(matrix effect) 또는 이온 억제(ion suppression)와 같은 분석 감도에 치명적인 영향을 미치는 문제를 일부 극복할 수 있으며 이에 따라 비교적 정확한 정량 결과를 얻을 수 있기에 일반적인 대사체 분석 과정은 컬럼을 이용한 성분 분리를 포함하는 것이 보통이다 [13]. 하지만 이러한 장점에도 불구하고 성분 분리 방법을 적용한 MS 의 결정적인 단점은 상대적으로 긴 분석시간(~10-60 min/시료)으로 인한 다수 시료 분석에 부적합하다는 것이다. 질량 분석기의 경우 온도, 습도 등 주변 환경에 커다란 영향을 받기에 시료수가 수백, 수천단위인 대규모 분석을 진행할 경우 긴 분석시간으로 인해 재현성을 보장하지 못하는 등의 어려움이 필연적으로 따른다 [14].

Direct-infusion MS (DI-MS)는 성분 분리법을 이용하지 않고 시료를 바로 질량분석기에 주입 및 분석하는 방법으로써 모든 대사체의 스팩트럼 정보를 시료 당 수 초~수십 초 이내로 확보할 수 있다 [15]. 물론 성분 분리를 진행하지 않기에 매트릭스 효과 및 이온 억제로 인한 데이터 손실과 이에 따른 정확한 정량의 불가능함과 같은 단점을 가지고 있으나, 사전 분석을 통해 시료의 대사체 정보를 확보한 상황에서 tandem mass spectrometry (MS/MS) 및 multiple reaction monitoring (MRM)과 같은 기법을 이용, 특정 대사체만을 선택적으로 선별하여 분석할 경우, LC-MS 및 GC-MS 를 이용한 수준의 높은 민감도, 선택성(selectivity) 및 정밀도(precision) 를 확보할 수 있다. 또한 fragmentation pattern 을 통해 해당 대사체를 구명하는 것 또한 가능하다 [16, 17]. 다시 말해서 DI-MS 는 기존의 LC-MS 및 GC-MS 를 대체할 수 있는 우수한 분석 속도, 민감도, 정확도, 선택성을 가진 high-throughput 분석 기법이라 할 수 있다 [14, 18].

DI-MS 와 같이 high-throughput 분석 기법의 개선에도 불구하고 판별을 위한 통계 도구(chemometric tool)의 개발 및 적용은, 특히 필연적으로 수백, 수천 점 이상의 다수 시료를

대상으로 하는 식품 또는 약용작물 분야에 있어 심각하게 정체되어 있다 [19]. 2017년에 발표된 대사체학에 따른 통계 도구 사용 빈도 조사에 따르면 대사체학 판별 분석을 위해 사용하는 도구 중 다변량 통계분석(multivariate statistical analysis)의 경우 principal component analysis (PCA, 96%) 및 partial least squares discriminant analysis (PLS-DA, 73%)가 여전히 압도적인 비율로 사용되고 있으며 가장 대중적인 기계 학습 기법 중 하나인 random forest (RF)의 경우 20% 대의 낮은 사용빈도를 보여주고 있다 [20]. 물론 PCA 및 PLS-DA 가 판별 분석을 위한 우수한 통계 도구임에는 의심의 여지가 없으나, PCA 는 본질적으로 판별 모델 제작에 적합한 방법이라고 하기는 어려우며, PLS-DA 의 경우 과적합(overfitting)문제 및 이상점(outlier)에 매우 취약하다는 심각한 단점을 가지고 있을 뿐 아니라 특히 혼입 시료와 같은 특정 대사체의 함량이 매우 유사한 시료를 대상으로 하는 판별 분석 시 결과를 보장할 수 없다 [21, 22]. 이에 따라 기존의 대사체 함량의 차이를 이용하는 다변량 통계분석과는 달리 대사체 변화 정도 및 변화 패턴을 인식하여 판별을 진행하는 기계 학습(machine learning)을 도입한 판별 분석이 대두되고 있다 [23, 24].

기계 학습은 인공 지능(artificial intelligent, AI)의 분야로써, 변수 간의 관계를 분석, 수학적인 공식을 이용하여 모델링을 하는 통계학적 모델링(statistical modelling)과는 다르게 특정 법칙 없이 입력한 데이터로부터 정보를 습득하여 특정한 모델을 구축하는 알고리즘을 말한다. 오믹스 분야에서의 기계 학습은 기존의 다변량 분석과 달리 노이즈를 포함한 다수의 변수를 전 처리 없이 동시에 처리가 가능하며, 모델의 과적합 및 이상점 문제에서 상대적으로 자유롭다 [22]. 지금까지는 유전체학(genomics)과 같은 바이오 인포메틱스(bio-informatiocs) 분야에서 특정 질병의 진단을 위한 유전자 마커 발굴을 위해 주로 사용되어 왔으나 유전자에 비해

훨씬 수가 많고 복잡한 대사체의 변화를 다루는 분야인 대사체학으로 그 범위가 확장되고 있다 [23, 25].

대사체학에서의 기계 학습은 단순한 판별 연구에서부터 활발히 사용되고 있는데, Maione et al 은 성공적으로 기계 학습 알고리즘 중 RF, support vector machine (SVM) 및 neural network (NNet)를 이용하여 쌀 내의 원소 20 종을 이용한 원산지 판별을 성공적으로 진행한 바 있으며 시료의 개수가 상대적으로 적고 독립 밸리데이션 연구가 시행되지 않았음에도 불구하고 우수한 판별 결과를 보여주고 있다 [26]. 해당 연구는 기계 학습이 다수의 시료 및 혼합된 시료의 판별에도 충분히 적용될 수 있음을 보여준다.

본 연구는 한국산, 중국산 쌀(Oryza Sativa L.) 및 이의 임의 혼합 시료(중국산 쌀 5%, 10%, 15%, 20%, 25%, 50%, 75%의 비율로 한국산 쌀과 혼입) 총 430 점 및 서로 다른 4 가지 품종의 인삼 시료(*P. ginseng, P. notoginseng, P. quinquefolius, P. vietnamensis*) 총 40 점을 대상으로 MS 기반, 특히 DI-MS 위주의 대사체학 판별 분석 전략을 설명한다. 구체적인 연구 내용은 1)LC-MS 를 이용한 untargeted metabolomics 를 진행, 쌀과 인삼의 판별 대사체 확보, 2)판별 대사체를 대상으로 DI-MS 기반의 신속 분석법을 적용하여 판별 대사체의 함량 정보를 얻음과 동시에 대사체 구조 동정 및 구명을 진행, 3)분석 결과를 다변량 통계분석인 PCA, PLS-DA 를 이용해 판별, 및 4)기계 학습 기법인 RF, SVM, C5.0 decision tree(C5.0), NNet, k-nearest neighbor (kNN)을 이용하여 판별 분석을 진행하고 다변량 통계분석과 기계 학습 결과를 비교하는 과정을 포함하고 있다. 연구 결과, DI-MS 를 이용하여 수백 점의 쌀과 인삼 시료를 수시간 이내로 신속하게 분석할 수 있었으며, 기계 학습 기반의 판별 분석이 binary classification 뿐 아니라 multiclass classification 에도 훌륭하게 적용될 수 있음을 확인하였다. 특히 판별이 매우 까다로운 혼입 시료의 판별에 있어 다변량 통계분석에 비해 기계 학습이 훨씬 우수한 결과를 보이는 것을 알 수 있었다.

2. Materials and Methods

2. 1. Materials and reagents

2. 1. 1. Rice

2014 년, 2015 년 및 2016 년에 재배된 한국산 및 중국산 쌀 각 80 점을 현지에 가서 직접 구매하였다. 구매 후 대사체의 변화를 막기 위해 −70 ° C 에서 보관한다. 시료 추출 및 기기 분석을 위한 용매로써 J. T. Baker 에서 구매한 HPLC grade acetonitrile, isopropanol, water 를 사용하였다. 스펙트럼을 보정하기 위한 내부표준품으로 Caffeine(Sigma−Aldrich, St. Louis, MO, USA)를 사용하였다. 마찬가지로 Sigma−Aldrich 에서 formic acid 를 구매하여 버퍼로써 용매에 첨가하였다. 추출 후 필터링을 위해 Polytetrafluoroethylene (PTFE) syringe filters with a pore size of 0.2 μm(Advantec, Tokyo, Japan)를 사용하였다.

본 연구에서는 한국산, 중국산 쌀 뿐 아니라 임의의 비율로 혼합한 쌀 또한 판별 대상으로써 제작, 분석하였다. 편의를 위해 한국산 쌀을 K, 중국산 쌀을 C 로 지칭하며 2014 년 재배된 쌀을 Batch A, 2015 년 재배된 쌀을 Batch B, 2016 년 재배된 쌀을 Batch C 로 지칭한다. 구체적인 쌀 시료의 정보는 Table 1 과 같다.

확보한 쌀 시료에 각각 고유한 숫자를 제공한 후 (K1, K2…. K30, C1, C2…. C30), Table 2 와 같이 R 로 자동 생성한 랜덤한 수열로 혼합 시료를 제작한다(한국산 쌀 3 종 및 중국산 쌀 3 종 혼합하여 혼합 시료 한 세트 제작). Batch A 및 Batch B 의 경우 0%, 25%, 50%, 75%, 100%의 총 5 가지 조건의 혼합 비율의 쌀 시료를 제작하며 Batch C 의 경우 0%, 5%, 10%, 15%, 20%, 25%의 6 가지 조합의 비율의 혼합 쌀 시료를 제작한다. 75%K 시료에 25%C 시료가 혼합된 경우 K/25%C 와 같이 표기한다. 최종적으로 Batch A 에서 150 점의 시료, Batch B 에서 100 점, Batch C 에서 180 점 총 430 점의 혼합 쌀 시료가 완성된다.

Table 1. Sample characteristics: White rice origins, cultivars.

재배 연도	2014 (Batch A)		2015 (Batch B)		2016 (Batch C)	
국가	원산지	수	원산지	수	원산지	수
한국	경기	8	경기	5	경기	8
	전남	8	전남	4	전남	8
	강원	4	강원	3	강원	4
	경북	4	경북	3	경북	4
	충남	2	충남	2	충남	2
	전북	2	전북	1	전북	2
	충북	1	충북	1	충북	1
	경남	1	경남	1	경남	1
중국	흑룡강	13	흑룡강	7	흑룡강	13
	랴오닝	8	랴오닝	6	랴오닝	8
	지린	6	지린	6	지린	6
	산둥	3	산둥	1	산둥	3
총계		60		40		60

Table 2. Sample characteristics: Random mixing sequences.

2014 (Batch A)		2015 (Batch B)		2016 (Batch C)	
M1	6 + 21 + 17	M1	11 + 6 + 8	M1	16 + 9 + 12
M2	6 + 28 + 27	M2	14 + 2 +	M2	21 + 3 + 7
M3	4 + 25 + 14	M3	6 + 20 + 12	M3	9 + 8 + 18
M4	17 + 30 + 7	M4	9 + 13 + 11	M4	13 + 19 + 16
M5	23 + 6 + 12	M5	3 + 12 + 7	M5	4 + 18 + 11
M6	26 + 29 + 7	M6	9 + 1 + 5	M6	13 + 2 + 8
M7	14 + 3 + 19	M7	8 + 16 + 19	M7	12 + 25 + 29
M8	12 + 25 + 5	M8	13 + 15 + 7	M8	19 + 23 + 10
M9	11 + 15 + 5	M9	9 + 14 + 16	M9	13 + 21 + 24
M10	11 + 28 + 4	M10	5 + 15 + 7	M10	8 + 23 + 10
M11	1 + 5 + 23	M11	11 + 2 + 4	M11	17 + 3 + 5
M12	27 + 15 + 18	M12	18 + 9 + 14	M12	27 + 13 + 21
M13	26 + 9 + 19	M13	17 + 19 + 13	M13	25 + 28 + 20
M14	5 + 29 + 9	M14	11 + 6 + 5	M14	16 + 8 + 7
M15	4 + 5 + 27	M15	1 + 14 + 5	M15	1 + 22 + 7
M16	24 + 29 + 10	M16	4 + 1 + 9	M16	5 + 1 + 14
M17	16 + 24 + 1	M17	3 + 16 + 7	M17	4 + 24 + 10
M18	1 + 20 + 27	M18	19 + 5 + 9	M18	29 + 8 + 14
M19	9 + 24 + 23	M19	4 + 12 + 9	M19	6 + 17 + 13
M20	30 + 18 + 20	M20	10 + 8 + 20	M20	15 + 12 + 30
M21	24 + 26 + 18			M21	1 + 4 + 14
M22	8 + 25 + 13			M22	12 + 25 + 2
M23	12 + 14 + 7			M23	18 + 13 + 3
M24	2 + 8 + 9			M24	7 + 3 + 15
M25	2 + 6 + 29			M25	20 + 16 + 2
M26	23 + 9 + 25			M26	17 + 11 + 27
M27	13 + 17 + 10			M27	8 + 7 + 25
M28	21 + 1 + 12			M28	14 + 7 + 18
M29	6 + 25 + 28			M29	7 + 1 + 23
M30	10 + 22 + 30			M30	9 + 5 + 29

2. 1. 2. Ginseng

각 10 점의 4 가지 서로 다른 품종의 인삼 시료(*P. ginseng, P. notoginseng, P. quinquefolius, P. vietnamensis)*를 한국 및 베트남 현지에서 직접 구매하였다. 용매의 버퍼 로써 formic acid 를 이용하였다 (Sigma Aldrich, St. Louis, MO, USA). 추출, 필터링 및 기기 분석을 위한 용매로써 J. T. Baker 에서 구매한 HPLC grade acetonitrile, methanol, water 를 사용하였다 필터링을 위해 Polytetrafluoroethylene (PTFE) syringe filters with a pore size of 0.2 μm (Advantec, Tokyo, Japan) 및 Oasis® SPE cartridges (Waters, Milford, MA, USA) 를 사용하였다.

2. 2. Sample preparation

2. 2. 1. Rice

동결건조, 분쇄 후 두개의 채(425 및 125 μm)를 이용해 걸러진 쌀 시료 150 mg 를 재현성을 확인하기 위한 내부표준품 caffeine 1mg 과 섞는다. 해당 혼합 시료에 6 mL 의 75% isopropanol 을 첨가한 후 100° C 의 워터 배스 에서 2 시간동안 추출한다. 이후 추출물을 5 분동안 12,000 rpm, 상온 조건에서 원심분리를 진행한 후 상징액을 0.2 μm PTFE filter 로 필터링 하여 이를 시료로 이용한다. 모든 시료의 분석 시퀀스는 바이어스를 최소화하기 위해 시료의 번호와 관계없이 랜덤하게 지정한다.

2. 2. 2. Ginseng

동결건조 후 분쇄 및 채(425 및 125 μm)를 이용해 걸러진 인삼 시료 20 mg 을 1 mL 100% methanol 을 이용해 추출한다. 추출은 상온에서 30 분동안 소니케이터에서 진행한다. 이후

추출물을 5 분동안 12,000 rpm, 상온 조건에서 원심분리 한 후 상징액을 0.2 μm PTFE filter 로 필터링 하여 깨끗한 vial 로 옮겨 담은 후 질소 퍼지로 용매를 제거, 이후 1 mL 의 water 에 재추출한다. 인삼 시료의 경우 추출물에 포함된 당과 같은 매트리스를 제거하기 위해 solid phase extraction (SPE)를 진행하며 순서 및 조건은 다음과 같다. 1)SPE 카트리지에 1 mL 의 methanol 및 water 를 순서대로 넣어 컨디셔닝을 진행한다. 2)1 mL 의 인삼 water 추출물을 카트리지에 로딩한 후 2 mL 의 20% methanol 을 넣어 카트리지를 씻어낸다. 3)1 mL 의 methanol 을 넣어 시료를 씻어내고 해당 추출물을 10 배 희석한 것을 시료로써 사용한다 [14]. 분석 시 쌀 시료와 마찬가지로 랜덤 시퀀스를 이용한다.

2. 3. LC-MS analysis condition

2. 3. 1. Rice

판별 대사체 선별을 위하여 Batch A, Batch B 에서 쌀 시료를 각각 20 점씩 선택해서 LC-MS 를 이용한 분석을 진행하였다. 쌀 시료의 LC-MS 분석은 Agilent triple-quadrupole MS 6460 system (Agilent, USA)을 이용하였다. 컬럼은 Acquity™ Ultra-performance liquid chromatography (UPLC) 용 column (BEH C18, 1.7 μm, 2.1 mm x 100 mm)을 이용하였다. 시료 주입 용량은 5 μL 로 설정하였다. 용매 A 는 water + 0.1% formic acid 이며 용매 B 는 acetonitrile + 0.1% formic acid 이다. 유속은 0.17 mL/min 이며 gradient 조건은 Table 3-(a)에 표기되어 있다. MS 분석은 electro spray ionization (ESI) ion source 이용하여 진행하였으며 분석 도중 lock mass 를 주입하여 분자량(*m/z*)의 흔들림을 보정하였다. MS 의 세부 조건은 Table 3-(b)와 같다.

Table 3. LC-gradient and mass spectometry condition for white rice analysis.

(a)

LC-gradient condition		
	용매 A (%)	용매 B (%)
0 min	100	0
5 min	70	30
15 min	30	70
25 min	25	75
27 min	0	100
50 min		

(b)

Mass spectrometry condition	
Ion mode	Negative
Scan time	200 scans/s
Accelerator	7 V
Fragmentor	135 V
Gas flow	11 L/min
m/z range	50-1000
Capillary voltage	4 kV

2. 3. 2. Ginseng

인삼 시료의 LC 기반 성분 분리를 위해 HPLC 2795 system (Waters Corporation, Milford, MA, USA)을 사용하였다. 컬럼은 Perkin-Elmer C18 column (50×2.1 mm, 1.9 μm, Norwalk, CT, USA)이다. 시료 주입 용량은 5 μL 이다. 사용한 용매 종류는 쌀 시료 분석에서 이용한 것과 동일하다. Gradient 조건은 Table 4-(a)와 같다. 유속은 0.17 mL/min 이며 컬럼 오븐 온도는 30° C 로 유지하였다. MS 분석은 ESI ion source 가 장비된 Micro triple quadrupole mass spectrometer (Waters Corporation, Milford, MA, USA)을 이용해 진행하였으며 구체적인 조건은 Table 4-(b)에 명시되어 있다.

Table 4. LC-gradient and mass spectometry condition for ginseng analysis.

(a)

LC-gradient condition		
	용매 A (%)	용매 B (%)
0 min	83	17
17 min	67	33
33 min	58	42
35 min	0	100
45 min	0	100

(b)

Mass spectrometry condition	
Ion mode	Negative
Gas flow	500 L/h
Source temp	110 ° C
m/z range	100-1500
Con voltage	90 V
Capillary voltage	-2 kV

2. 4. DI-MS analysis condition

2. 4. 1. Rice

쌀 시료의 분석을 위해 우선 product ion mode (MS/MS)를 이용하여 타겟 대사체의 구조 동정 및 구명을 진행하고 MRM transition 을 설정한다. 이후 MRM mode 를 이용하여 타겟 대사체만을 선별적으로 정밀하게 분석한다. 분석을 위한 MS 조건은 Table 3-(b)에 기술한 조건과 동일하나 최적의 대사체 검출을 위해 Negative ion mode 만이 아닌 positive ion mode 도 함께 사용하였으며 collision energy 를 20 eV 로 설정하였다. 최적의 감도를 확보하기 위하여 lysophosphatidlycholine (lysoPC)의 분석은 positive ion mode 를 이용, lysophosphatidylethanolamine (lysoPE) 및 lysophosphatidlyglycerol (lysoPG)의 분석을 위해서는 negative ion mode 를 사용한다. 분석 도중 ion source 의 오염을 방지하기 위해 50%의 용매 A 를 0.2 mL/min 의 유속으로 계속해서 흘려준다. 시료에 대한 피크는 0.25 초에서 나타나지만 결과의 재현성 및 오염 방지를 위해 분석 시간은 30 초로 설정하였다.

2. 4. 2. Ginseng

인삼 시료의 경우 daughter ion scan mode (MS/MS)를 적용해 타겟 대사체의 구조 동정, 구명을 진행한다. 쌀 시료 분석과 마찬가지로 MS 조건은 Table 4-(b)와 같은 동일한 조건에서 타겟 대사체 별로 다른 collision energy 를 적용한다(*m/z* 1107, 945 = 40 eV, *m/z* 1149, 783 = 30 eV). 인삼 시료의 경우 오토샘플러를 이용하지 않고 오토 실린지 펌프를 통해 ion source 로 바로 시료를 주입한다. 시료 주입 속도는 10 μL/min 로 설정하였다.

2. 5. Data treatment

2. 5. 1. Rice

모든 LC-MS 결과 데이터는 .mzdata 포멧으로 저장한 후 MZmine 2.23.을 이용해 전처리를 진행한다 [27]. DI-MS 결과 데이터의 경우 Agilent Mass Hunter Workstation software version B.06.00 프로그램을 이용, 스팩트럼의 intensity 정보를 직접 수집한다. 대사체의 동정은 METLIN metabolite database (http://metlin.scripps.edu/) 및 In-laboratory library 를 이용하였다 [28].

2. 4. 2. Ginseng

모든 인삼 LC-MS 분석 결과 데이터는 쌀 시료와 동일한 방법으로 전 처리한다. DI-MS 결과의 경우 MassLynx program 을 이용, 스팩트럼의 intensity 정보를 직접 수집한다. 대사체의 동정은 In-laboratory library 를 이용하여 진행하였다.

2. 6. Data processing and visualization

해당 연구의 전체적인 과정은 Figure 1 을 통해 확인할 수 있다. 다변량 통계분석인 PCA 및 PLS-DA 는 Metaboanalyst 3.0 를 이용하여 진행한다 [29]. DI-MS 분석을 위한 타겟 대사체 선별은 variable importance in projection (VIP) score 를 기준으로 1 이상인 대사체를 대상으로 한다. 모든 LC-MS 및 DI-MS 를 통해 얻어진 데이터는 PCA, PLS-DA 를 이용하여 판별 분석하기 전에 pareto-scaling algorithm 을 이용해 data centering 및 scaling 을 진행한다.

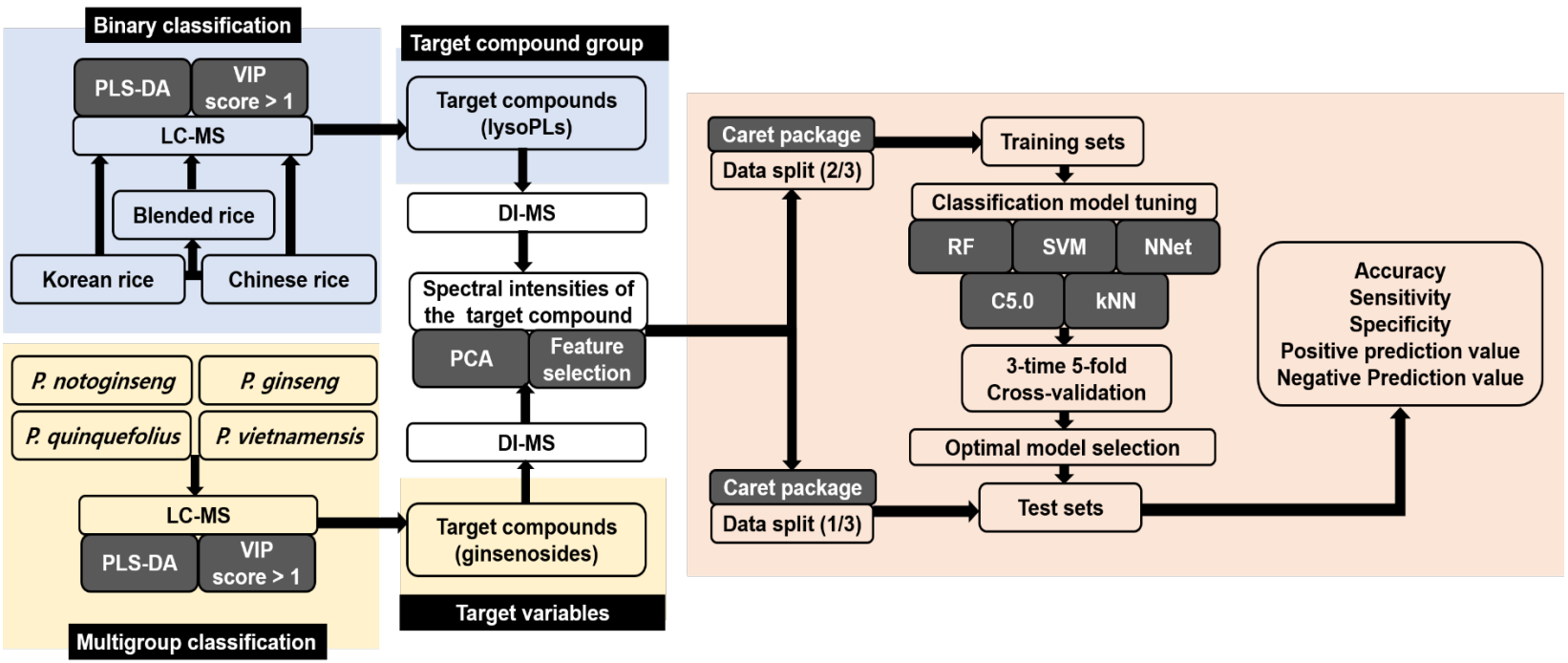


Figure 2. The workflow of data collection and analysis.

2. 7. Machine learning and discriminant analysis

기계 학습 알고리즘을 데이터에 적용하기 전에 Boruta R package version 5.2.0 기반의 wrapper algorithm 을 이용하여 유의미한 대사체 선별 작업을 진행한다 [30]. Wrapper algorithm 은 쌀 시료의 분석 데이터에만 적용하며 선별 알고리즘을 적용하지 않은 인삼 시료 분석 데이터 결과와 차이를 비교하였다. 이후 선별된 대사체 정보를 기계 학습을 이용한 판별 분석에 이용한다. 다변량 통계분석을 위한 데이터 처리와 마찬가지로 판별 분석 이전에 Yeo-Johnson transformation 기반의 data centering 및 scaling을 선행한다. 분석 결과의 67%를 training set, 33%를 test set 으로 설정하였다. RF, SVM, C5.0, NNet 및 kN 의 총 5 가지 기계 학습 알고리즘을 사용하였으며 모든 판별 모델을 제작하기 전에 최적화 과정을 진행하여 최적화된 패러미터를 판별 모델에 적용하였다. 기계 학습 알고리즘 별 최적 패러미터는 다음과 같다. RF=randomly chosen variables (mtry); SVM=kernel smoothing parameter (sigma) 및 cost (C); C5.0=trials, model 및 winnow; NNet=size, decay 및 bag; kNN=the number of closest training examples (k) [31, 32]. 한국산 쌀을 positive group 으로 설정하고 중국산 쌀 및 혼입 쌀은 negative group 으로 설정하였다. 설정된 최적의 패러미터는 3-time repeated 5-fold cross-validation 을 이용한 resampling procedure 를 통해 다시 한번 최적화 과정을 거쳤다. Training set 을 이용한 판별 모델 제작 후 test set 을 적용하여 판별 모델의 정확도(accuracy), 민감도(sensitivity), 특이도(specificity), 양성예측도(positive predictive value, PPV) 및 음성예측도(negative predictive value, NPV)를 평가하였다. 모든 판별 분석은 caret R package version 6.0-73 을 이용해 진행하였다 [33].

3. Result and discussion

3. 1. Multivariate statistical analysis of the LC-MS results

3. 1. 1. Rice

각각 20 점의 한국산 쌀 시료 및 중국산 쌀 시료를 대상으로 LC-MS 분석을 진행한 후 타겟 대사체 검출을 위해 PLS-DA 를 이용해 판별 모델을 제작하였다. 이후 cross-validation 및 1000-time permutation test 를 진행하여 과적합된 판별 모델이 아님을 검증한 후 VIP score 를 이용하여 판별 대사체를 검출하였다. Figure 2 의 PLS-DA 플롯과 같이 우수한 판별 결과를 확인할 수 있다. Cross-validation 결과인 goodness of fit (R^2), goodness of prediction (Q^2) 수치 및 permutation test 결과 또한 해당 PLS-

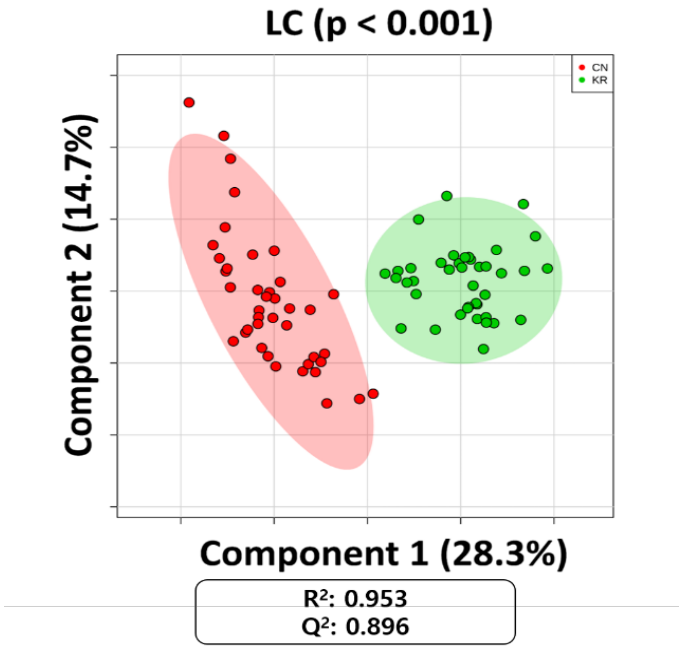


Figure 2. PLS-DA result of LC-MS analysis for white rice.

DA 모델이 과적합이 일어나지 않은, 우수한 판별 모델임을 증명한다.

해당 PLS-DA 결과에서 타겟 대사체를 확인하기 위하여 VIP score 가 1 이상인 대사체만을 선택적으로 확인하였다 (Table 5). 흥미롭게도 In-laboratory library 및 METLIN 을 이용한 동정 결과 모든 판별 대사체가 lysoPC, lysoPE, lysoPG 와 같은 lysophospholipids (lysoPLs) 대사체로 드러났다.

Table 5. Identified discrimination marker of white rice.

성분 명	Mass per charge ratio (m/z)			VIP
	측정 값	실제 값	Adduct ion	score
lysoPE(18:3)	474.261	474.262	$[M-H]^-$	1.777
lysoPE(18:2)	476.277	476.278	$[M-H]^-$	1.435
lysoPE(18:1)	478.294	478.293	$[M-H]^-$	1.478
lysoPC(14:0)	512.300	512.299	$[M-H+HCOOH]^-$	3.454
lysoPC(16:1)	538.314	538.315	$[M-H+HCOOH]^-$	1.416
lysoPC(18:2)	564.332	564.330	$[M-H+HCOOH]^-$	1.411
lysoPC(16:0)	540.330	540.330	$[M-H+HCOOH]^-$	1.167
lysoPG(16:0)	483.271	483.272	$[M-H]^-$	2.464

3. 1. 2. Ginseng

각 10 점의 인삼 시료의 LC-MS 분석 결과를 토대로 PLS-DA 를 이용해 판별 모델을 제작하였다. 마찬가지로 cross-validation 및 1000-time permutation test 를 진행하여 과적합된 판별 모델이 아닌지 검증한 후, VIP score 를 이용하여 판별 대사체를 확인하였다. 판별 결과 Figure 3 과 같이 4 종의 서로 다른 인삼 품종이 깔끔하게 판별된 것을 확인할 수 있다.

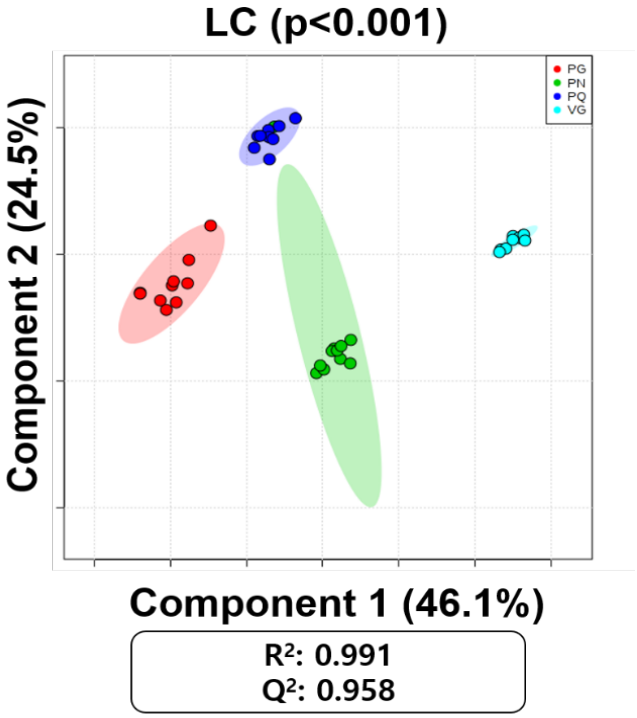


Figure 3. PLS-DA result of LC-MS analysis for ginseng.

Table 6 는 VIP score 가 1 이상인 대사체를 정리한 것이다. Waters MS 를 이용한 분석의 경우 감도가 Agilent MS 에 비해 상대적으로 매우 떨어지므로 측정 값은 소수점 첫째자리까지 기록하였다. 4 종의 타겟 대사체는 모두 ginsenoside 로 확인되었다.

Table 6. Identified discrimination marker of ginseng.

성분 명	Mass per charge ratio (*m/z*)		VIP
	측정 값	Adduct ion	score
Ginsenoside Rg2	783.5	$[M-H]^-$	3.257
Ginsenoside Re	945.5	$[M-H]^-$	2.456
Ginsenoside Rb1	1107.5	$[M-H]^-$	2.895
Quinquenoside R1	1149.2	$[M-H]^-$	1.040

3. 2. Target compounds characterization

3. 2. 1. Rice

LC-MS 결과를 통해 쌀 내의 lysoPL 가 주요 판별 대사체가 될 수 있음을 확인하였다. 타겟 대사체로 확인된 lysoPL 은 Phospholipid (PLs) 중 fatty acid 가 하나로 구성된 PL 을 말하며 쌀 내의 lysoPLs 는 쌀의 전분 성분, 특히 amylose 와 결합하여 starch-lipid 라는 특수한 구조를 형성, 쌀의 특징 및 품질을 결정하는 매우 중요한 성분으로 알려져 있다 [34]. 우선 DI-MS 분석에 앞서 타겟 대사체들의 daughter ion 정보를 확보, 타겟 대사체 구조 동정과 동시에 MRM transition 을 선택하였다. 일반적으로 lysoPC 는 positive ion mode 에서 높은 감도를 보여주며 lysoPC 를 제외한 나머지 lysoPLs 의 경우 negative ion mode 에서 높은 감도를 나타내는 것으로 알려져 있기에 product ion scan 시 lysoPC 는 positive, lysoPE, lysoPG 는 negative ion mode 를 이용하였다[17]. Figure 4 를 확인하였을 때 lysoPC 의 경우 *m/z* 184.1, $[phosphocholine+H]^+$ ion 이 탄소 개수 및 이중결합의 수와 관계없이 가장 intensity 가 높은 daughter ion 으로 나타나는 것을 확인 할 수 있다. 반면 lysoPE 및 lysoPG 의 경우 해당 성분군만의 특징적인 daughter ion (lysoPE=*m/z* 196.1, lysoPG= *m/z* 153.1) 이외에도 결합된 고유한 fatty acid 의 spectrum 이 가장 intensity 가 높은 daughter ion 으로 나타나는 것을 확인할 수 있다. 최종적으로 각 대사체의 product ion scan mode 에서 확인된 가장 intensity 가 높은 daughter ion 은 DI-MS 를 이용한 분석 시 precursor ion 과 함께 MRM transition 을 형성한다.

재미있게도 인체 유래 시료와는 달리 쌀에 존재하는 지방산은 myristic acid (C14:0, 0.2%), palmitic acid (C16:0, 15.6%), palmitoleic acid (C16:1, 0.2%), stearic acid (C18:0, 1.4%), oleic acid (C18:1, 39.4%), linoleic acid (C18:2, 40.6%),

linolenic acid (C18:3, 1.5%)의 비율로 존재하며 특정 탄소 개수 이상의 지방산은 포함되어 있지 않다. 이에 따라 쌀에 존재하는 lysoPLs 또한 특정 탄소 및 이중결합으로만 이루어져 있다 [35]. 즉 쌀 시료에서 찾아낸 판별 대사체인 lysoPLs 의 경우 그 수가 수십개 내외로 한정적인 만큼, 범위를 확장 시켜 쌀에 존재하는 모든 lysoPLs 의 동시 스크리닝 방법 개발을 진행하였다. lysoPLs 는 lysoPC, lysoPE, lysoPG 이외에도 lysophosphatidlyinositol (lysoPI), lysophosphathydylserine (lysoPS) 및 lysophosphatidic acid (lysoPA)의 subgroup 을 가지고 있다. 모든 lysoPLs 의 쌀 함유 여부를 확인하기 위하여 product ion scan 를 통한 fragmentation pattern 을 확인함으로써 예상되는 adduct 의 분자량을 통한 검출 결과를 통해 총 17 종의 타겟 lysoPLs 를 최종적으로 확정하고 각각의 MRM transition 을 설정하였다 (Table 7~10).

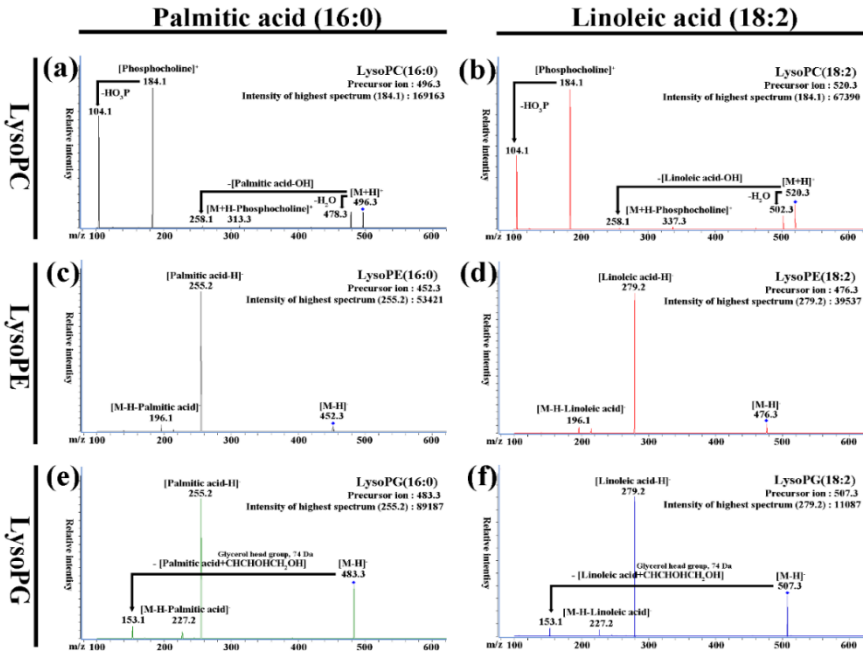


Figure 4. The MS/MS fragmentation and structure elucidation of six presentative lysoPCs.

Table 7. The *m/z* of [fatty acid−H]$^-$ ion in negative ion mode.

	14:0 (myristic acid)	16:0 (palmitic acid)	16:1 (palmitoleic acid)	18:0 (stearic acid)	18:1 (oleic acid)	18:2 (linoleic acid)	18:3 (linolenic acid)
[fatty acid−H]$^-$	227.2	255.2	253.2	283.2	281.2	279.2	277.2

Table 8. The *m/z* of lysophospholipid precursors.

	Ion type	14:0	16:0	16:1	18:0	18:1	18:2	18:3
lysoPC	[M+H]$^+$	468.3	496.3	494.3	524.3	522.3	520.3	518.3
lysoPE	[M−H]$^-$	424.3	452.3	450.3	480.3	478.3	476.3	474.3
	[M−H+HCOOH]$^-$	470.3	498.3	496.3	526.3	524.3	522.3	520.3
lysoPG	[M−H]$^-$	455.3	483.3	481.3	511.3	509.3	507.3	505.3
	[M−H+HCOOH]$^-$	501.3	529.3	527.3	557.3	555.3	553.3	551.3
lysoPI	[M−H]$^-$	543.3	571.3	569.3	599.3	597.3	595.3	593.3
	[M−H+HCOOH]$^-$	589.3	617.3	615.3	645.3	643.3	641.3	639.3
lysoPS	[M−H]$^-$	468.3	496.3	494.3	524.3	522.3	520.3	518.3
	[M−H+HCOOH]$^-$	514.3	542.3	540.3	570.3	568.3	566.3	564.3
lysoPA	[M−H]$^-$	381.3	409.3	407.3	437.3	435.3	433.3	431.3
	[M−H+HCOOH]$^-$	427.3	455.3	453.3	483.3	481.3	479.3	477.3

Table 9. The list of identified lysophospholipids.

	Ion type	14:0	16:0	16:1	18:0	18:1	18:2	18:3
lysoPC	[M+H]$^+$	O	O	O	O	O	O	Δ
lysoPE	[M−H]$^-$	O	O	O	O	O	O	O
	[M−H+HCOOH]$^-$	X	X	X	X	X	X	X
lysoPG	[M−H]$^-$	O	O	Δ	X	O	O	X
	[M−H+HCOOH]$^-$	X	X	X	X	X	X	X
lysoPI	[M−H]$^-$	X	X	X	X	X	X	X
	[M−H+HCOOH]$^-$	X	X	X	X	X	X	X
lysoPS	[M−H]$^-$	X	X	X	X	X	X	X
	[M−H+HCOOH]$^-$	X	X	X	X	X	X	X
lysoPA	[M−H]$^-$	X	X	X	X	X	X	X
	[M−H+HCOOH]$^-$	X	X	X	X	X	X	X

Table 10. The characteristics of 17 targeted phospholipid species.

Ion mode	성분 명	Precursor ion type	Precursor ion (*m/z*)	Product ion type	Product ion (*m/z*)
Positive	lysoPC(14:0)	$[M+H]^+$	468.3	$[Phosphocholine]^+$	184.1
	lysoPC(16:0)	$[M+H]^+$	496.3	$[Phosphocholine]^+$	184.1
	lysoPC(16:1)	$[M+H]^+$	494.3	$[Phosphocholine]^+$	184.1
	lysoPC(18:0)	$[M+H]^+$	524.3	$[Phosphocholine]^+$	184.1
	lysoPC(18:1)	$[M+H]^+$	522.3	$[Phosphocholine]^+$	184.1
	lysoPC(18:2)	$[M+H]^+$	520.3	$[Phosphocholine]^+$	184.1
Negative	lysoPE(14:0)	$[M-H]^-$	424.3	$[Myristic\ acid-H]^-$	227.2
	lysoPE(16:0)	$[M-H]^-$	452.3	$[Palmitic\ acid-H]^-$	255.2
	lysoPE(16:1)	$[M-H]^-$	450.3	$[Palmitoleic\ acid-H]^-$	253.2
	lysoPE(18:0)	$[M-H]^-$	480.3	$[Stearic\ acid-H]^-$	283.2
	lysoPE(18:1)	$[M-H]^-$	478.3	$[Oleic\ acid-H]^-$	281.2
	lysoPE(18:2)	$[M-H]^-$	476.3	$[Linoleic\ acid-H]^-$	279.2
	lysoPE(18:3)	$[M-H]^-$	474.3	$[Linolenic\ acid-H]^-$	277.2
	lysoPG(14:0)	$[M-H]^-$	455.3	$[Myristic\ acid-H]^-$	227.2
	lysoPG(16:0)	$[M-H]^-$	483.3	$[Palmitic\ acid-H]^-$	255.2
	lysoPG(18:1)	$[M-H]^-$	509.3	$[Oleic\ acid-H]^-$	281.2
	lysoPG(18:2)	$[M-H]^-$	507.3	$[Linoleic\ acid-H]^-$	279.2

3. 2. 2. Ginseng

LC-MS 분석 결과, 총 4 종의 ginsenoside 가 판별 대사체로 확인 되었다. 하지만 인삼의 경우 쌀의 타겟 대사체인 lysoPLs 와 같이 대사체 별로 precursor ion 의 분자량과 daughter ion 의 분자량이 확실히 구분되는 것과는 달리, 특정 ginsenoside 의 daughter ion 이 다른 ginsenoside 의 precursor 의 분자량이 될 수 있다. Table 11 와 Figure 5 은 ginsenoside 의 이러한 특징을 잘 보여준다. 예를 들어 quinquenoside R1 의 precursor ion 인 *m/z* 1149.2 를 타겟으로 product ion scan 을 할 경우 ginsenoside Re, 또는 ginsenoside Rd 의 precursor ion 과 분자량이 동일한 *m/z* 945, 및 ginsenoside Rb1 의 precursor ion 의 분자량과 같은 *m/z* 1107.5 가 생성되는 것을 확인할 수 있다. 이러할 경우 DI-MS 에서 *m/z* 1107.5 를 타겟으로 분석된 특정 스펙트럼은

quinquenoside R1 의 daughter ion 의 함량과 ginsenoside Rb1 의 precursor ion 의 함량을 동시에 포함하게 된다. LC 와 같은 성분 분리를 적용하여 분석하는 경우에는 ginsenoside 의 종류에 따라 머무름 시간의 차이가 발생하므로 물질 동정에 큰 무리가 없으나 DI-MS 를 이용할 경우 성분 분리가 진행되지 않으므로 특정 스팩트럼이 어떠한 ginsenoside 로부터 유래된 것인지 확인하기 어려워지는 문제점이 있다.

다시 말해 DI-MS 에서 특정 분자량을 타겟으로 선정할 경우 분석 결과로 도출된 스팩트럼이 다양한 ginsenoside 의 precursor ion 과 daughter ion 를 동시에 포함하게 된다. 즉, 쌀의 lysoPLs 과는 달리 단일 스팩트럼을 특정 성분이라 지칭할 수 없게 되는 것이다.

Table 11. Fragmentation patterns of the four target ions from ginseng.

Target ion	Fragment	Fragmentation	Target ion	Fragment	Fragmentation
1107	459.3	aglycone for PPD type	783	459.3	aglycone for PPD type
	621.4	783-Glc		621.4	783-Glc
	764.7	783-water		637.4	783-xyl,Ara
	783.5	945-Glc		783.5	fragment from others
	945.5	1107-Glc			
	1107.5	Rb1+ fragment from others			
945	459.3	aglycone for PPD type	1149	459.3	aglycone for PPD type
	475.3	aglycone for PPT type		621.4	783-Glc
	621.4	783-Glc		765.5	783-water
	637.4	783-Glc		783.5	945-Glc
	765.5	783-water		926.3	945-water
	783.5	945-Glc		945.5	1107-Glc
	945.5	parent+ fragment from		1088.5	1107-water
				1107.5	1149- (Acetyl-H)
				1149.2	Quinqenoside R1

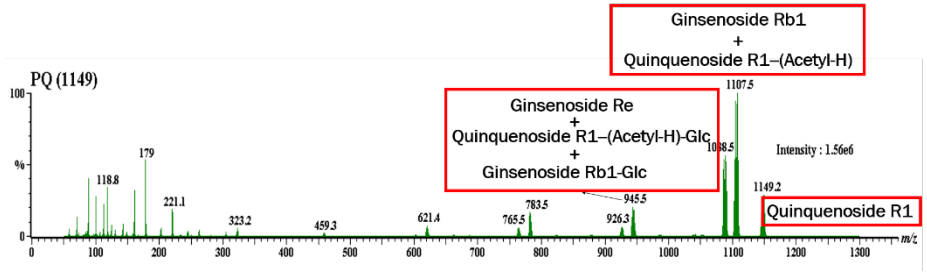


Figure 5. Fragmentation pattern of *m/z* 1149 on DI-MS analysis.

따라서 ginsenoside 와 같은 대사체의 경우 판별을 위해 특정 성분을 판별 대사체로 지정하지 않고 판별 대사체의 precursor ion 및 이의 daughter ion 의 *m/z* 값 만을 판별 대사체로써 이용 해야만 한다. 실제로 판별 만을 목적으로 할 경우 굳이 판별 대사체를 구명하지 않아도 *m/z* 정보만을 이용하여 판별을 진행하는 것이 가능하다.

3. 3. DI-MS analysis and data exploration-box plot with PCA

3. 3. 1. Rice

17 종의 lysoPLs 를 대상으로 총 430 점의 쌀 시료를 분석하였다. 첫번째로 Batch A 의 5 개 그룹(100%K, 100%C, 및 혼합 쌀 K/25%C, K/50%C, K/75%C) 결과를 이용, PCA 를 이용하여 스코어 플롯을 제작하였으며 각 대사체의 intensity 기반으로 박스 플롯을 제작하였다. Figure 6-(a)를 살펴보면 100%K 및 100%C 의 경우 비교적 깔끔하게 판별이 일어나는 것을 확인할 수 있다. 하지만 혼합 쌀 시료의 경우 사실상 판별이 불가능할 정도로 그룹이 중첩되어 있음을 확인할 수 있다. Batch B(Figure 6-(b)) 또한 Batch A 와 마찬가지의 결과를 보여주고 있다.

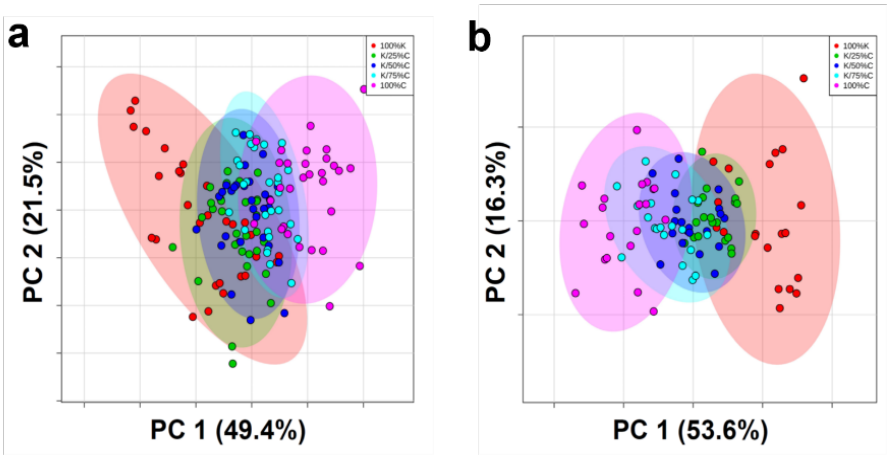


Figure 6. PCA analysis of white rice DI-MS result: (a)Batch A and (b)Batch B.

박스 플롯(Figure 7)을 통해서도 마찬가지의 결과를 확인 할 수 있다. 100%K 및 100%C 의 경우 비교적 명확하게 대사체의 농도 차이가 드러나지만 혼합 쌀의 대사체 농도를 통해서는 판별이 불가능에 가깝다. 특이한 점은 대사체의 변화 경향성이 Batch A(Figure 7-(a)), Batch B(Figure 7-(b))에서 크게 다르지 않다는 점이다. 전반적으로 한국 쌀에서는 lysoPC 농도가 높게 나타나고 있으며 중국 쌀에서는 lysoPE 가 높은 농도를 보여주고 있다. 즉 재배 년도는 lysoPLs 의 농도에 상대적으로 큰 영향을 미치지 않으며 재배 환경이 대사체의 농도, 특히 lysoPLs 의 농도에 영향을 미친다는 것을 박스 플롯 결과를 통해서 예측 할 수 있다.

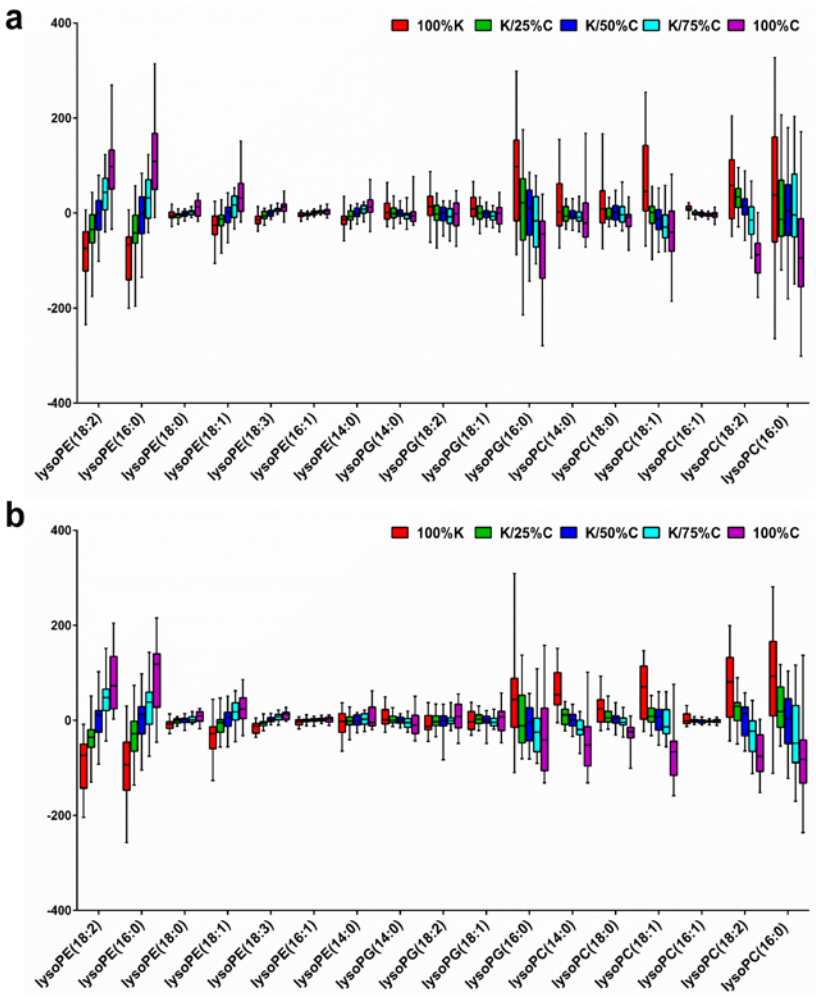


Figure 7. Box plots of the relative differences in 17 lysoPL concentrations.

3. 3. 2. Ginseng

4 종의 서로 다른 인삼 (*P. ginseng*;PG, *P. notoginseng*;PN, *P.quinquefolius*;PQ, *P.vietnamensis*;VG), 을 대상으로 확보한 4 종의 판별 대사체를 타겟으로 하여 precursor ion 및 daughter ion 의 스팩트럼 정보를 확보하였다. 인삼 사료의 경우 single ion monitoring(SIM)을 이용해 분석한 precursor ion 만 판별 대사체로 사용할 때 및 daughter ion scan mode 를 이용하여 daughter ion 까지 판별 대사체로 포함 시켰을 때의 결과를 비교하였다. Figure 8-(a)는 precursor ion 만 판별 대사체로 사용하였을 때의 PCA plot 이며 Figure 8-(b)는 precursor ion 및 이의 daughter ion 까지 판별 대사체로 포함 시켰을 때의 PCA 플롯 이다. 두 경우 모두 4 종의 서로 다른 인삼의 판별 정도가 우수하나 특히 daughter ion 을 포함하였을 때의 결과가 그렇지 않은 경우에 비해 *P. notoginseng* 및 *P. vietnamensis* 간에 좀 더 우수한 판별이 일어나는 것을 확인할 수 있다.

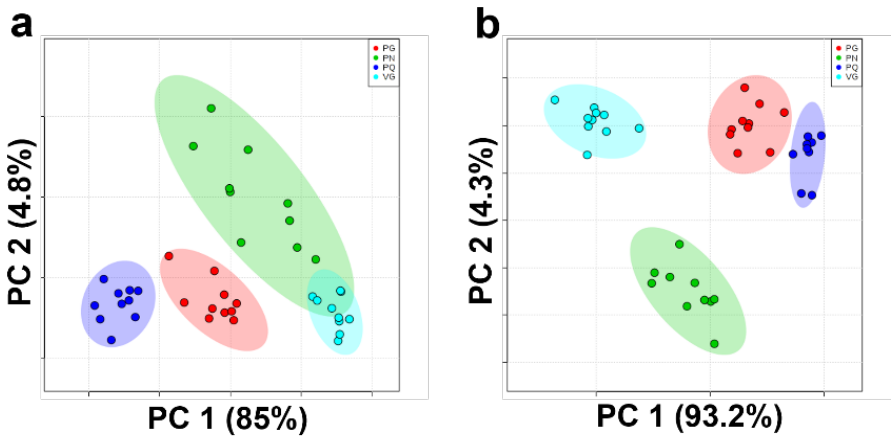


Figure 8. PCA analysis of ginseng DI-MS result: (a)four precursor ions and (b)precursor ions with their daughter ions.

박스 플롯 결과(Figure 9)를 살펴보면 인삼의 종류에 따라 daughter ion 의 생성 비율이 서로 다른 것을 확인할 수 있다. 인삼별로 서로 다른 판별 대사체의 함량 차이가 daughter ion 의 생성 비율 차이를 발생시키고 daughter ion 의 함량 차이가 판별을 위한 새로운 변수로 작용하여 더욱 우수한 판별 결과를 발생시키는 것으로 예측된다.

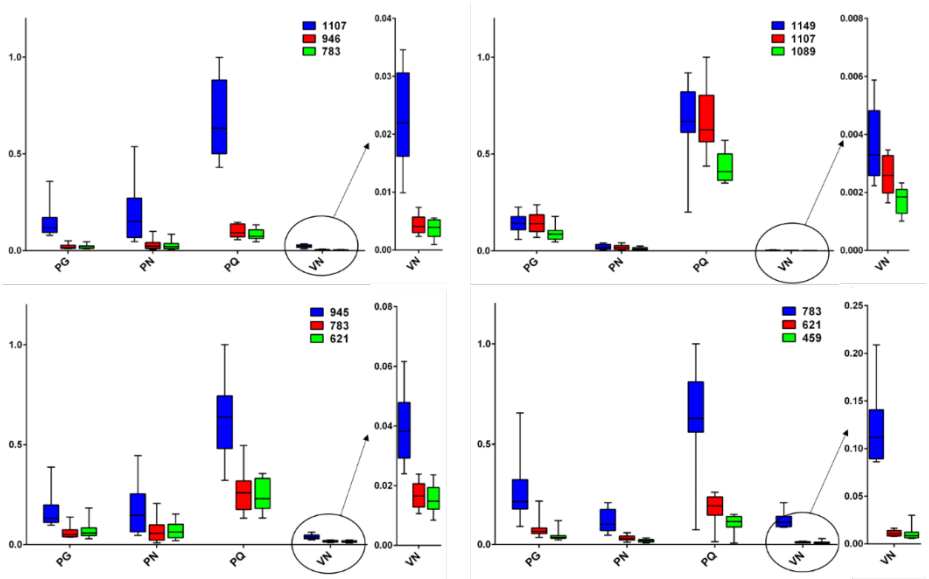


Figure 9. Box plots of the four discrimination precursor ions and their daughter ions.

3. 4. Machine learning

3. 4. 1. Rice

Batch A, Batch B 의 분석 결과를 RF, SVM, C5.0, NNet, kNN 의 총 5 종의 기계 학습 판별 기법을 적용하여 판별을 시도하였다. Training set 을 이용하여 판별 모델을 제작하였으며 test set 을

통해 해당 판별 모델의 우수성을 검증하였다. 타겟 대사체 정보를 이용하여 판별 모델을 제작하기 전 Batch A 의 100%K 및 100%C 분석 결과를 토대로 Boruta wrapper algorithm 을 통해 판별 모델 제작에 유의미한 대사체를 검출 하였다 (Figure 10).

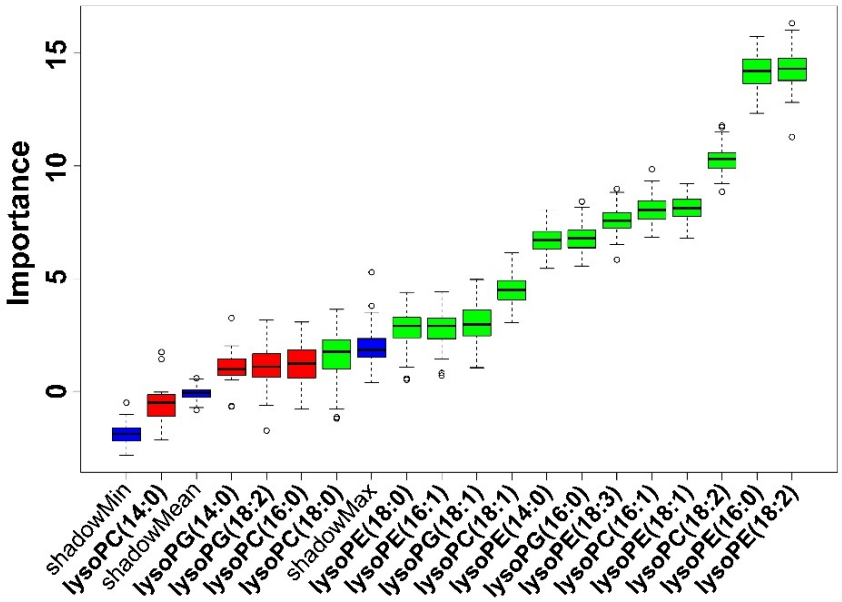


Figure 10. Feature selection for the white rice discrimination experiment.

17 종의 lysoPLs 중 13 종의 lysoPLs 가 판별 모델 제작에 의미 있는 대사체로 확인되었으며 lysoPC(16:0), lysoPC(14:0), lysoPG(18:2) 및 lysoPG(14:0)는 판별 모델 제작에서 제외되었다. Batch A 및 Batch B 를 이용한 최적의 판별 모델의 패러미터는 Table 12 에 표기되어 있다.

Table 12. The tuning parameters of the selected discrimination models for 2014 and 2015 white rice.

Discrimination	RF		SVM		C5.0		NNet		kNN	
Batch A (2014 white rice)										
K vs K/25%C										
	mtry	11	*Sigma*	0.13	*Model*	Rules	*Size*	16	*k*	11
			C	3.07	*Trials*	5	*Decay*	3.9E−5		
					Winnow	False	*Bag*	True		
K vs K/50%C										
	mtry	5	*Sigma*	0.13	*Model*	Rules	*Size*	16	*k*	11
			C	3.07	*Trials*	5	*Decay*	3.9E−5		
					Winnow	False	*Bag*	True		
K vs K/75%C										
	mtry	1	*Sigma*	0.02	*Model*	Rules	*Size*	2	*k*	7
			C	0.20	*Trials*	5	*Decay*	0.38		
					Winnow	False	*Bag*	True		
K vs C										
	mtry	1	*Sigma*	0.01	*Model*	Rules	*Size*	1	*k*	7
			C	0.20	*Trials*	5	*Decay*	0.29		
					Winnow	False	*Bag*	False		
Batch B (2015 white rice)										
K vs K/25%C										
	mtry	1	*Sigma*	0.13	*Model*	Rules	*Size*	14	*k*	7
			C	35.38	*Trials*	18	*Decay*	4.0E-5		
					Winnow	False	*Bag*	True		
K vs K/50%C										
	mtry	1	*Sigma*	0.01	*Model*	Rules	*Size*	1	*k*	5
			C	4.13	*Trials*	18	*Decay*	0.016		
					Winnow	False	*Bag*	True		
K vs K/75%C										
	mtry	1	*Sigma*	0.03	*Model*	Rules	*Size*	1	*k*	9
			C	1.08	*Trials*	18	*Decay*	0.02		
					Winnow	False	*Bag*	True		
K vs C										
	mtry	1	*Sigma*	0.01	*Model*	Rules	*Size*	1	*k*	9
			C	0.10	*Trials*	4	*Decay*	0.02		
					Winnow	True	*Bag*	True		

해당 패러미터를 이용해 판별 모델을 제작, test set 을 적용한 결과는 아래와 같다 (Table 13)

Table 13. The predictive performance of different classifiers for the test sets of 2014 and 2015 white rice.

Experiment	Method	Accuracy	Sensitivity	Specificity	PPV	NPV
Batch A (2014 white rice)						
K vs K/25%C	RF	0.75	0.80	0.70	0.73	0.78
	SVM	0.85	0.7	1.00	1.00	0.77
	C5.0	0.70	0.60	0.80	0.75	0.67
	NNet	0.90	1.00	0.80	0.83	1.00
	kNN	0.75	1.00	0.50	0.67	1.00
K vs K/50%C	RF	0.95	1.00	0.90	0.91	1.00
	SVM	1.00	1.00	1.00	1.00	1.00
	C5.0	0.90	1.00	0.80	0.83	1.00
	NNet	0.95	0.90	1.00	1.00	0.91
	kNN	0.95	1.00	0.90	0.91	1.00
K vs K/75%C	RF	1.00	1.00	1.00	1.00	1.00
	SVM	0.95	0.90	1.00	1.00	0.91
	C5.0	0.95	0.90	1.00	1.00	0.91
	NNet	1.00	1.00	1.00	1.00	1.00
	kNN	1.00	1.00	1.00	1.00	1.00
K vs C	RF	1.00	1.00	1.00	1.00	1.00
	SVM	1.00	1.00	1.00	1.00	1.00
	C5.0	0.95	0.90	1.00	1.00	0.91
	NNet	1.00	1.00	1.00	1.00	1.00
	kNN	1.00	1.00	1.00	1.00	1.00
Batch B (2015 white rice)						
K vs K/25%C	RF	0.92	1.00	0.83	0.86	1.00
	SVM	0.92	1.00	0.83	0.86	1.00
	C5.0	0.83	1.00	0.67	0.75	1.00
	NNet	0.83	1.00	0.67	0.75	1.00
	kNN	0.67	1.00	0.33	0.60	1.00
K vs K/50%C	RF	1.00	1.00	1.00	1.00	1.00
	SVM	1.00	1.00	1.00	1.00	1.00
	C5.0	0.92	1.00	0.83	0.86	1.00
	NNet	0.92	1.00	0.83	0.86	1.00
	kNN	0.83	1.00	0.67	0.75	1.00
K vs K/75%C	RF	1.00	1.00	1.00	1.00	1.00
	SVM	1.00	1.00	1.00	1.00	1.00
	C5.0	1.00	1.00	1.00	1.00	1.00
	NNet	1.00	1.00	1.00	1.00	1.00
	kNN	0.92	1.00	0.83	0.86	1.00
K vs C	RF	1.00	1.00	1.00	1.00	1.00
	SVM	1.00	1.00	1.00	1.00	1.00
	C5.0	0.92	0.83	1.00	1.00	0.86
	NNet	1.00	1.00	1.00	1.00	1.00
	kNN	1.00	1.00	1.00	1.00	1.00

기계 학습을 통한 판별 모델 제작 및 평가 결과를 살펴보면, 다변량 통계 분석 결과에서는 그룹간의 심각한 중첩으로 인해 사실상 판별 불가능하던 100%K 와 K/25%C 간의 혼합 쌀 간의 판별 정확도가 최소 70% 수준으로 상대적으로 매우 우수한 판별 결과를 보여준다. 100%K 와 K/50%C, K/75%C 을 판별할 경우 어떠한 판별 알고리즘을 사용하더라도 90% 이상의 매우 우수한 판별 결과를 보여주는 것을 알 수 있다. 전반적인 모든 모델 평가 결과를 고려하였을 때 Batch A 의 경우 NNet 이 가장 우수하고, RF, SVM, C5.0 은 평범, kNN 이 가장 우수하지 못한 판별 알고리즘으로 확인된다. Batch B 또한 Batch A 와 상당히 유사한 판별 결과를 보여준다. 하지만 Batch B 에서는 RF, SVM 가 가장 우수한 판별 알고리즘으로 확인되며 kNN 은 Batch A 와 마찬가지로 가장 우수하지 못한 알고리즘으로 확인되었다. 종합하자면 기계 학습을 이용한 판별 모델 제작을 통해 혼입 쌀과 같은 기존의 PCA, PLS-DA 와 같은 다변량 통계 분석을 이용하여 제작한 판별 모델로는 사실상 판별이 불가능한 그룹 또한 판별이 가능하다고 할 수 있다.

또한 기계 학습 방법이 어느 정도까지 판별이 가능한지 확인하기 위해 훨씬 적은 비율로 혼합된 쌀 시료군인 Batch C (K/5%C, K/10%C, K/15%C, K/20%C, K/25%C)를 이용, 기계 학습의 우수한 판별 능력을 증명 하였다. Table 14 는 Batch C 를 이용한 판별 모델의 tuning 패러미터이다.

Table 14. The tuning parameters of the selected discrimination models of 2016 white rice.

Discrimination	RF		SVM		C5.0		NNet		kNN	
K vs K/5%C										
	mtry	6	*Sigma*	0.12	*Model*	Rules	*Size*	3	*k*	1
			C	3.07	*Trials*	52	*Decay*	3.7E-5		
					Winnow	False	*Bag*	True		
K vs K/10%C										
	mtry	1	*Sigma*	0.12	*Model*	Rules	*Size*	9	*k*	5
			C	3.07	*Trials*	52	*Decay*	2.1E-3		
					Winnow	False	*Bag*	True		
K vs K/15%C										
	mtry	2	*Sigma*	0.13	*Model*	Rules	*Size*	9	*k*	5
			C	3.07	*Trials*	52	*Decay*	2.1E-3		
					Winnow	False	*Bag*	True		
K vs K/20%C										
	mtry	1	*Sigma*	0.01	*Model*	Rules	*Size*	1	*k*	5
			C	6.65	*Trials*	52	*Decay*	3.9E-3		
					Winnow	False	*Bag*	True		
K vs K/25%C										
	mtry	2	*Sigma*	6.5E-3	*Model*	Rules	*Size*	3	*k*	11
			C	9.45	*Trials*	5	*Decay*	3.7E-5		
					Winnow	False	*Bag*	True		

Batch C 의 test set 을 적용한 판별 결과는 Table 15 에서 확인할 수 있다. 재미있게도 Batch C 의 100%K vs K/25%C 결과는 Batch A, Batch B 의 100%K vs K/25%C 결과와 상당히 유사한 결과를 보여준다. Batch C 에서도 RF 와 SVM 이 가장 우수한 판별 알고리즘으로 확인되었으며 해당 알고리즘은 K/5%C 와 같은 미량이 혼입된 시료조차 판별 가능할 정도의 우수한 판별 모델을 제공한다. kNN 은 K vs K/5%C 에서는 우수한 결과를 보여주나 다른 그룹 간의 판별 결과는 좋지 못하며 NNet 및 C5.0 은 모든 그룹에서 좋지 않은 결과를 보인다. 한가지 흥미로운 결과는 NNet 및 kNN 의 negative prediction value 가 모두 1.00 을 나타내고 있다는 것으로 이는 단순히 중국 쌀의 혼합 여부만을 판단하는 판별 모델로서는 100% 의 정확도를 제공한다는 것을 의미한다.

Table 15. The predictive performance of different classifiers for the test sets of various mixing ratios of 2016 white rice.

Experiment	Method	Accuracy	Sensitivity	Specificity	PPV	NPV
K vs K/5%C	RF	0.80	0.80	0.80	0.80	0.80
	SVM	0.90	1.00	0.80	0.83	1.00
	C5.0	0.65	0.80	0.50	0.62	0.71
	NNet	0.70	1.00	0.40	0.63	1.00
	kNN	0.85	1.00	0.70	0.77	1.00
K vs K/10%C	RF	0.85	0.90	0.80	0.82	0.89
	SVM	0.90	1.00	0.80	0.83	1.00
	C5.0	0.75	0.80	0.70	0.73	0.78
	NNet	0.75	1.00	0.50	0.67	1.00
	kNN	0.70	1.00	0.40	0.63	1.00
K vs K/15%C	RF	0.90	1.00	0.80	0.83	1.00
	SVM	0.90	1.00	0.80	0.83	1.00
	C5.0	0.75	0.80	0.70	0.73	0.78
	NNet	0.80	1.00	0.60	0.71	1.00
	kNN	0.80	1.00	0.60	0.71	1.00
K vs K/20%C	RF	0.90	1.00	0.80	0.83	1.00
	SVM	0.90	1.00	0.80	0.83	1.00
	C5.0	0.80	0.90	0.70	0.75	0.88
	NNet	0.75	1.00	0.50	0.67	1.00
	kNN	0.85	1.00	0.70	0.77	1.00
K vs K/25%C	RF	0.90	1.00	0.80	0.83	1.00
	SVM	0.90	1.00	0.80	0.83	1.00
	C5.0	0.80	0.70	0.90	0.88	0.75
	NNet	0.85	1.00	0.70	0.77	1.00
	kNN	0.80	1.00	0.60	0.71	1.00

3 개의 batch 를 모두 고려하였을 때 RF 및 SVM 알고리즘이 가장 우수한 판별 알고리즘으로 확인 되었다.

3. 4. 2. Ginsenoside

인삼 시료의 경우 총 4 개의 그룹의 판별을 동시에 진행해야 하며 그룹 별 샘플의 수가 10 개로 쌀에 비해 매우 적다. 또한 쌀 시료와 달리 boruta wrapper algorithm 을 이용하여 판별에 의미 있는 변수를 따로 선별 하지도 않았기에 판별 모델을 제작하기에는 상대적으로 좋지 않은 조건을 가정하였다고 할 수 있다. 다수 그룹을 기계 학습을 통해 동시에 판별할 경우 판별 모델을 평가하기 위한 기준으로서 1 대 1 로 판별하는 경우와 달리 민감도, 특이도, 양성예측도 및 음성예측도는 기준으로 사용하기에 부적합하므로 정확도만을 기계 학습 알고리즘의 우수성을 검증하기 위한 척도로써 사용하였다. 4 종의 서로 다른 인삼 시료를 판별하기 위한 최적의 패러미터는 Table 16 과 같다.

Table 16. The tuning parameters of the selected discrimination models of four ginseng cultivars.

RF		SVM		C5.0		NNet		kNN	
mtry	26	Sigma	1.02	*Model*	Rules	*Size*	3	k	1
		C	23.268	*Trials*	5	*Decay*	4.76E-3		
				Winnow	True	*Bag*	True		

해당 패러미터를 통해 제작된 모델에 test set 을 도입하여 정확도를 계산한 결과 RF=0.98, SVM=0.64, C5.0=0.84, NNet=0.90, kNN=0.97 로 다수의 그룹, 적은 수의 데이터를 이용한 판별 모델 제작에서도 기계 학습은 매우 뛰어난 판별 결과를 보여준다. 이 중 RF 및 kNN 이 특히 우수한 판별 알고리즘으로 드러난다. 특히 RF 는 쌀, 인삼을 포함한 모든 판별 분석 연구에서 공통적으로 가장 우수한 기계 학습 알고리즘으로 확인 되었다.

3. 5. Random forest

판별을 위한 최적의 알고리즘으로 선정된 RF 는 결정 트리(decision tree)를 이용하는 대표적인 기계 학습 알고리즘 중 하나이다. 결정 트리는 시각적이고 명시적인 방법으로 특정 결과를 낸 과정을 보여주기에 결과를 해석하고 이해하기 매우 용이하며, 자료를 가공할 필요가 거의 없다. 하지만 데이터가 지나치게 복잡하고 특성을 구분하기 어려울 때 분류율이 대폭 감소하고 결정 트리가 복잡해지는 문제가 발생한다. 이는 NNet 등의 다른 기계 학습 알고리즘이 여러 변수를 동시에 고려하는 반면 결정 트리는 한 개의 변수만을 선택 가능하기 때문이다. 반면 RF 는 결정 트리의 이러한 단점을 극복하기 위해 훈련 과정에서 결정 트리를 수백, 수천 개 임의로 제작, 각 결정 트리 별 결과값을 평균화 하여 결과를 예측하게 된다. 이렇게 다수의 결정 트리를 이용할 경우, 기존의 단일 결정 트리의 장점인 데이터 전처리 과정의 불필요함과 더불어 높은 정확성, 과적합 문제로부터의 자유로움, 빠른 학습, 테스트 속도 및 RF 의 가장 중요한 특성 중 하나인 변수의 중요성에 순위를 매기는 것이 가능하게 된다 [36-38]. 이러한 RF 의 특성은 다수의 대사체 정보 처리를 요구하는 대사체학 기반의 연구에서 특히 유용하게 사용될 수 있다 [39-41]. 무엇보다 본 연구에서 데이터 확보를 위해 사용한 DI-MS 와 같은 분석 기법은 LC-MS 와 다르게 대사체의 머무름 시간 정보를 제공하지 않으며, 상대적으로 낮은 감도로 인해 데이터의 형태가 매우 복잡한 경우가 많은데 RF 는 이러한 데이터 처리에 가장 적합한 기계 학습 알고리즘이 될 수 있다.

4. Conclusion

본 연구는 쌀 시료(단일 그룹 및 혼합 시료 판별) 및 인삼 시료(다수 그룹 간의 동시 판별)의 대사체학 기반, DI-MS 분석 및 기계 학습을 적용한 판별을 진행하고 기존의 LC-MS 및 다변량 통계 분석 기반의 판별 연구와 비교하였다. 그 결과 DI-MS 는 LC-MS 에 비해 수십 배 이상의 빠른 분석 속도와 결과의 우수한 재현성을 보여주었으며 기계 학습을 통해 제작한 판별 모델은 기존의 PCA, PLS-DA 와 같은 다변량 통계 분석 보다 훨씬 우수한 판별 결과를 제공하였다. 본 연구에서 이용한 5 개의 기계 학습 알고리즘 중 Random forest 가 특히 우수한 알고리즘으로 확인되었다. 또한 기계 학습의 특성 상 더 많은 분석 결과를 투입할수록 더욱 더 우수한 결과를 보여주기에 High-throughtput 분석법인 DI-MS 분석 기법은 특히 기계 학습과 함께 응용할 수 있는 매우 유용한 분석 기법이라 할 수 있을 것이다.

5. References

[1] A. Alonso, S. Marsal, A. Julià, Analytical methods in untargeted metabolomics: state of the art in 2015, Frontiers in bioengineering and biotechnology, 3 (2015).

[2] R. Díaz, O.J. Pozo, J.V. Sancho, F. Hernández, Metabolomic approaches for orange origin discrimination by ultra−high performance liquid chromatography coupled to quadrupole time−of−flight mass spectrometry, Food chemistry, 157 (2014) 84−93.

[3] Y. Jung, J. Lee, J. Kwon, K.−S. Lee, D.H. Ryu, G.−S. Hwang, Discrimination of the geographical origin of beef by 1H NMR−based metabolomics, Journal of agricultural and food chemistry, 58 (2010) 10458−10466.

[4] M.−Y. Choi, W. Choi, J.H. Park, J. Lim, S.W. Kwon, Determination of coffee origins by integrated metabolomic approach of combining multiple analytical data, Food Chemistry, 121 (2010) 1260−1268.

[5] D.I. Ellis, H. Muhamadali, D.P. Allen, C.T. Elliott, R. Goodacre, A flavour of omics approaches for the detection of food fraud, Current Opinion in Food Science, 10 (2016) 7−15.

[6] M. Beckmann, D.P. Enot, D.P. Overy, J. Draper, Representation, comparison, and interpretation of metabolome fingerprint data for total composition analysis and quality trait investigation in potato cultivars, Journal of Agricultural and Food Chemistry, 55 (2007) 3444−3451.

[7] R.C. De Vos, S. Moco, A. Lommen, J.J. Keurentjes, R.J. Bino, R.D. Hall, Untargeted large−scale plant metabolomics using liquid chromatography coupled to mass spectrometry, Nature protocols, 2 (2007) 778−791.

[8] S. Tiziani, Y. Kang, J.S. Choi, W. Roberts, G. Paternostro, Metabolomic high−content nuclear magnetic resonance−based drug screening of a library of kinase inhibitors, Nature communications, 2 (2011) 545.

[9] J.L. Markley, R. Brüschweiler, A.S. Edison, H.R. Eghbalnia, R. Powers, D. Raftery, D.S. Wishart, The future of NMR−based metabolomics, Current opinion in biotechnology, 43 (2017) 34−40.

[10] S. Mas, S.G. Villas-Bôas, M. Edberg Hansen, M. Åkesson, J. Nielsen, A comparison of direct infusion MS and GC-MS for metabolic footprinting of yeast mutants, Biotechnology and Bioengineering, 96 (2007) 1014−1022.

[11] J. Trygg, J. Gullberg, A.I. Johansson, P. Jonsson, H. Antti, S.L. Marklund, T. Moritz, Extraction and GC/MS analysis of the human blood plasma metabolome, Analytical Chemistry, 77 (2005) 8086−8094.

[12] D.K. Lim, C. Mo, J.H. Lee, N.P. Long, Z. Dong, J. Li, J. Lim, S.W. Kwon, The integration of multiplatform MS-based metabolomics and multivariate analysis for the geographical origin discrimination of Oryza sativa L, Journal of Food and Drug Analysis, DOI (2017).
[13] T.M. Annesley, Ion suppression in mass spectrometry, Clinical chemistry, 49 (2003) 1041-1044.
[14] S. Kim, B.K. Shin, D.K. Lim, T.J. Yang, J. Lim, J.H. Park, S.W. Kwon, Expeditious discrimination of four species of the Panax genus using direct infusion-MS/MS combined with multivariate statistical analysis, J Chromatogr B Analyt Technol Biomed Life Sci, 1002 (2015) 329-336.
[15] L. Lin, Q. Yu, X. Yan, W. Hang, J. Zheng, J. Xing, B. Huang, Direct infusion mass spectrometry or liquid chromatography mass spectrometry for human metabonomics? A serum metabonomic study of kidney cancer, Analyst, 135 (2010) 2970-2978.
[16] D.K. Lim, N.P. Long, C. Mo, Z. Dong, L. Cui, G. Kim, S.W. Kwon, Combination of mass spectrometry-based targeted lipidomics and supervised machine learning algorithms in detecting adulterated admixtures of white rice, Food Research International, 100 (2017) 814-821.
[17] D.K. Lim, C. Mo, N.P. Long, G. Kim, S.W. Kwon, Simultaneous Profiling of Lysoglycerophospholipids in Rice (Oryza sativa L.) Using Direct Infusion-Tandem Mass Spectrometry with Multiple Reaction Monitoring, J Agric Food Chem, 65 (2017) 2628-2634.
[18] D.K. Lim, C. Mo, N.P. Long, J. Lim, S.W. Kwon, A rapid and reliable method for discriminating rice products from different regions using MCX-based solid-phase extraction and DI-MS/MS-based metabolomics approach, Journal of Chromatography B, 1061 (2017) 185-192.
[19] S.P. Putri, Y. Nakayama, F. Matsuda, T. Uchikata, S. Kobayashi, A. Matsubara, E. Fukusaki, Current metabolomics: practical applications, Journal of bioscience and bioengineering, 115 (2013) 579-589.
[20] R.J. Weber, T.N. Lawson, R.M. Salek, T.M. Ebbels, R.C. Glen, R. Goodacre, J.L. Griffin, K. Haug, A. Koulman, P. Moreno, Computational tools and workflows in metabolomics: An international survey highlights the opportunity for harmonisation through Galaxy, Metabolomics, 13 (2017) 12.
[21] B. Xi, H. Gu, H. Baniasadi, D. Raftery, Statistical analysis and modeling of mass spectrometry-based metabolomics data, Mass spectrometry in metabolomics: methods and protocols, DOI (2014) 333-353.
[22] P.S. Gromski, H. Muhamadali, D.I. Ellis, Y. Xu, E. Correa, M.L. Turner, R. Goodacre, A tutorial review: Metabolomics and partial least

squares–discriminant analysis–a marriage of convenience or a shotgun wedding, Analytica chimica acta, 879 (2015) 10–23.
[23] R. Goodacre, S. Vaidyanathan, W.B. Dunn, G.G. Harrigan, D.B. Kell, Metabolomics by numbers: acquiring and understanding global metabolite data, Trends in biotechnology, 22 (2004) 245–252.
[24] N.P. Long, D.K. Lim, C. Mo, G. Kim, S.W. Kwon, Development and assessment of a lysophospholipid–based deep learning model to discriminate geographical origins of white rice, Scientific Reports, 7 (2017) 8552.
[25] J. Taylor, R.D. King, T. Altmann, O. Fiehn, Application of metabolomics to plant genotype discrimination using statistics and machine learning, Bioinformatics, 18 (2002) S241–S248.
[26] C. Maione, B.L. Batista, A.D. Campiglia, F. Barbosa, R.M. Barbosa, Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry, Computers and Electronics in Agriculture, 121 (2016) 101–107.
[27] T. Pluskal, S. Castillo, A. Villar–Briones, M. Orešič, MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry–based molecular profile data, BMC bioinformatics, 11 (2010) 395.
[28] C.A. Smith, G. O'Maille, E.J. Want, C. Qin, S.A. Trauger, T.R. Brandon, D.E. Custodio, R. Abagyan, G. Siuzdak, METLIN: a metabolite mass spectral database, Therapeutic drug monitoring, 27 (2005) 747–751.
[29] J. Xia, I.V. Sinelnikov, B. Han, D.S. Wishart, MetaboAnalyst 3.0—making metabolomics more meaningful, Nucleic acids research, 43 (2015) W251–W257.
[30] M.B. Kursa, W.R. Rudnicki, Feature selection with the Boruta package: Journal, DOI (2010).
[31] M. Fernández–Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems, J. Mach. Learn. Res, 15 (2014) 3133–3181.
[32] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P. Feuston, Random forest: a classification and regression tool for compound classification and QSAR modeling, Journal of chemical information and computer sciences, 43 (2003) 1947–1958.
[33] M. Kuhn, Caret package, Journal of Statistical Software, 28 (2008) 1–26.
[34] L. Liu, D.L. Waters, T.J. Rose, J. Bao, G.J. King, Phospholipids in rice: significance in grain quality and health benefits: a review, Food chemistry, 139 (2013) 1133–1145.

[35] T.J. Rose, L. Liu, M. Wissuwa, Improving phosphorus efficiency in cereal crops: is breeding for reduced grain phosphorus concentration part of the solution?, Frontiers in plant science, 4 (2013).
[36] D.R. Cutler, T.C. Edwards, K.H. Beard, A. Cutler, K.T. Hess, J. Gibson, J.J. Lawler, Random forests for classification in ecology, Ecology, 88 (2007) 2783-2792.
[37] G. Biau, Analysis of a random forests model, Journal of Machine Learning Research, 13 (2012) 1063-1095.
[38] A. Hapfelmeier, T. Hothorn, K. Ulm, C. Strobl, A new variable importance measure for random forests with missing data, Statistics and Computing, 24 (2014) 21-34.
[39] T. Chen, Y. Cao, Y. Zhang, J. Liu, Y. Bao, C. Wang, W. Jia, A. Zhao, Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection, Evidence-Based Complementary and Alternative Medicine, 2013 (2013).
[40] C. Lanz, A.D. Patterson, J. Slavík, K.W. Krausz, M. Ledermann, F.J. Gonzalez, J.R. Idle, Radiation metabolomics. 3. Biomarker discovery in the urine of gamma-irradiated rats using a simplified metabolomics protocol of gas chromatography-mass spectrometry combined with random forests machine learning algorithm, Radiation research, 172 (2009) 198-212.
[41] L. Zhou, Q. Wang, P. Yin, W. Xing, Z. Wu, S. Chen, X. Lu, Y. Zhang, X. Lin, G. Xu, Serum metabolomics reveals the deregulation of fatty acids metabolism in hepatocellular carcinoma and chronic liver diseases, Analytical and bioanalytical chemistry, 403 (2012) 203-213.
[42] Z. Zhou, K. Robards, S. Helliwell, C. Blanchard, Composition and functional properties of rice, International journal of food science & technology, 37 (2002) 849-868.
[43] C. Maningat, B. Juliano, Starch lipids and their effect on rice starch properties, Starch-Stärke, 32 (1980) 76-82.
[44] Ö. Tokuşoğlu, C.A. Hall III, Fruit and cereal bioactives: sources, chemistry, and applications, CRC Press2011.
[45] M. Godet, V. Tran, M. Delage, A. Buléon, Molecular modelling of the specific interactions involved in the amylose complexation by fatty acids, International Journal of Biological Macromolecules, 15 (1993) 11-16.
[46] H. Yamakawa, M. Hakata, Atlas of rice grain filling-related metabolism under high temperature: joint analysis of metabolome and transcriptome demonstrated inhibition of starch accumulation and induction of amino acid accumulation, Plant and Cell Physiology, 51 (2010) 795-809.

[47] N. Ahmed, I.J. Tetlow, S. Nawaz, A. Iqbal, M. Mubin, N. ul Rehman, M. Shah, A. Butt, D.A. Lightfoot, M. Maekawa, Effect of high temperature on grain filling period, yield, amylose content and activity of starch biosynthesis enzymes in endosperm of basmati rice, Journal of the Science of Food and Agriculture, 95 (2015) 2237-2243.
[48] K. Kawaminami, M. Kojima, M. Ohnishi, S. Ito, Comparative composition of brown rice lipids (lipid fractions) of indica and japonica rices, Bioscience, biotechnology, and biochemistry, 63 (1999) 619-626.
[49] M. Kleinschmidt, V.A. McMahon, Effect of growth temperature on the lipid composition of Cyanidium caldarium, Plant physiology, 46 (1970) 286-289.
[50] S.H. Cho, T.M. Cheesbrough, Warm growth temperatures decrease soybean cholinephosphotransferase activity, Plant physiology, 93 (1990) 72-76.

Acknowledgements

본 연구에서 사용된 내용, Table 및 Figure 는 본인이 주저자 및 공저자로 참여한 문헌 1) Lim, Dong Kyu, et al. "Combination of mass spectrometry-based targeted lipidomics and supervised machine learning algorithms in detecting adulterated admixtures of white rice." *Food Research International* 100 (2017): 814-821, 2) Long, Nguyen Phuoc, et al. "Development and assessment of a lysophospholipid-based deep learning model to discriminate geographical origins of white rice." *Scientific Reports* 7.1 (2017): 8552. 3) Lim, Dong Kyu, et al. "The integration of multiplatform MS-based metabolomics and multivariate analysis for the geographical origin discrimination of Oryza sativa L." *Journal of Food and Drug Analysis* (2017) 4) Lim, Dong Kyu, et al. "Simultaneous Profiling of Lysoglycerophospholipids in Rice (Oryza sativa L.) Using Direct Infusion-Tandem Mass Spectrometry with Multiple Reaction Monitoring." *Journal of Agricultural and Food Chemistry* 65.12 (2017): 2628-2634 및 5) Kim, Shinae, et al. "Expeditious discrimination of four species of the Panax genus using direct infusion-MS/MS combined with multivariate statistical analysis." *Journal of Chromatography B* 1002 (2015): 329-336. 에 사용된 것을 수정, 개선한 것이다.

Abstract

In this study, we suggest Direct Infusion-Mass Spectrometry (DI-MS) and machine learning based metabomics strategy. This strategy presents that are significantly improved in terms of analysis time and discrimination accuracy, compared with conventional methods that mainly apply column chromatography and multivariate statistical analysis. To prove the applicability of the proposed strategy, two differently originated rice (Korean rice, Chinese rice, and rice mixed at an arbitrary ratio) and four different cultivars of ginseng (*P. ginseng, P. notoginseng, P. quinquefolius, and P. Vietnamensis*) were analyzed and discriminated using different analytical devices (rice: Agilent triple quadrupole 6460 system, ginseng: WatersMirco triple quadrupole mass spectrometer). The marker metabolites for the classification were revealed as lysophospholipids of rice and ginsenosides of ginseng by liquid chromatography-mass spectrometry (LC-MS) result. After that, the marker metabolites were simultaneously analyzed using a DI-MS-based approach, which provided an expeditious analysis time of about 30 seconds per sample and excellent sensitivity. In order to construct discriminant models based on the acquired data from DI-MS approach, multivariate statistical analysis (principal component analysis and partial least squares discriminant analysis) and machine learning classifiers (random forest, support vector machine, C5.0, neural network, and k-nearest neighbor) were applied and evaluated, respectively. As a result of discriminant analysis, especially, two kinds of rice and four kinds of ginseng were discriminated more accurately when using machine learning classifiers. In particular, random forest showed highly precise discrimination results to discriminate mixed rice samples mixed into about 5% which cannot be determined by multivariate statistical analysis. In conclusion, DI-MS and machine learning based metabolomics strategy for the discriminant analysis is highly suitable method for experiments requiring large samples, quick speed, and excellent discrimination accuracy.

Keywords: Discriminant analysis, Direct infusion-mass spectrometry, Machine learning

Supporting information

Table S1. Detailed information of korean white rice cultivars

재배 년도	2014 (Batch A)		2015 (Batch B)		2016 (Batch C)	
Korea	Label	Cultivar	Label	Cultivar	Label	Cultivar
	KR1	Choochung	KR1	Choochung	KR1	Ode
	KR2	Ode	KR2	Samgwang	KR2	Hopyeong
	KR3	Samgwang	KR3	Samgwang	KR3	Samgwang
	KR4	Senoori	KR4	Shindongjin	KR4	Samgwang
	KR5	Hopyeong	KR5	Choochung	KR5	Shindongjin
	KR6	Hitomebore	KR6	Ode	KR6	Ilpum
	KR7	Ilpum	KR7	Ode	KR7	Shindongjin
	KR8	Shindongjin	KR8	Hopyeong	KR8	Choochung
	KR9	Shindongjin	KR9	Shindongjin	KR9	Jinsang
	KR10	Ode	KR10	Ilmi	KR10	Ode
	KR11	Samgwang	KR11	Samgwang	KR11	Samgwang
	KR12	Ode	KR12	Ilpum	KR12	Samgwang
	KR13	Samgwang	KR13	Choochung	KR13	Senoori
	KR14	Samgwang	KR14	Ilmi	KR14	Ilmi
	KR15	Jinsang	KR15	Ilmi	KR15	Hitomebore
	KR16	Koshihikari	KR16	Jinsang	KR16	Koshihikari
	KR17	Choochung	KR17	Hitomebore	KR17	Choochung
	KR18	Choochung	KR18	Koshihikari	KR18	Ode
	KR19	Samgwang	KR19	Choochung	KR19	Ode
	KR20	Choochung	KR20	Ode	KR20	Ode
	KR21	Ode			KR21	Choochung
	KR22	Ode			KR22	Choochung
	KR23	Ilmi			KR23	Ilmi
	KR24	Ilmi			KR24	Samgwang
	KR25	Samgwang			KR25	Choochung
	KR26	Shindongjin			KR26	Samgwang
	KR27	Ode			KR27	Shindongjin
	KR28	Ilmi			KR28	Ilmi
	KR29	Ilmi			KR29	Ilmi
	KR30	Choochung			KR30	Ode

Table S2. Detailed information of chinese white rice cultivars

재배 년도	2014 (Batch A)		2015 (Batch B)		2016 (Batch C)	
China	Label	Cultivar	Label	Cultivar	Label	Cultivar
	CN1	Dongbeidami	CN1	Dongbeidami	CN1	Dongbeidami
	CN2	Zhenzhumi	CN2	Zhanglixiangmi	CN2	Dongbeidami
	CN3	Zhenzhumi	CN3	Baijinxiangmi	CN3	Youjidami
	CN4	Dongbeidami	CN4	Zhenzhumi	CN4	Zhenzhumi
	CN5	Daohuaxiang	CN5	Youjidami	CN5	Baijinxiangmi
	CN6	Xuejingdao	CN6	Yalujiang 7 xi	CN6	Zhenzhumi
	CN7	Zhanglixiangmi	CN7	Dongbeidami	CN7	Dongbeidami
	CN8	Wuchangdami	CN8	Shengtaidao	CN8	Daohuaxiang
	CN9	Youjidami	CN9	Yatianmi	CN9	Daohuaxiang
	CN10	Daohuaxiang	CN10	Lusedami	CN10	Yalujiang 7 xi
	CN11	Fuxiangdao	CN11	Xuejingdao	CN11	Wuchangdami
	CN12	Jinjingdao	CN12	Fuxiangdao	CN12	Yueguangdaoxi
	CN13	Dongbeidami	CN13	Daohuaxiang	CN13	Jinjingdao
	CN14	Zhanglixiangmi	CN14	Zhanglixiangmi	CN14	Xuejingdao
	CN15	Yatianmi	CN15	Zhenzhumi	CN15	Daohuaxiang
	CN16	Daohuaxiang	CN16	Wuchangdami	CN16	Lusedami
	CN17	Youjidami	CN17	Zhanglixiangmi	CN17	Zhanglixiangmi
	CN18	Youjidami	CN18	Daohuaxiang	CN18	Youjidami
	CN19	Baijinxiangmi	CN19	Youjidami	CN19	Youjidami
	CN20	Wuchangxiangmi	CN20	Daohuaxiang	CN20	Yatianmi
	CN21	Dongbeidami			CN21	Yalujiang 3 xi
	CN22	Lusedami			CN22	Fuxiangdao
	CN23	Shengtaidao			CN23	Shengtaidao
	CN24	Zhonghuahemi			CN24	Zhonghuahemi
	CN25	Yueguangdaoxi			CN25	Daohuaxiang
	CN26	Zhanglixiangmi			CN26	Zhanglixiangmi
	CN27	Yalujiang 7 xi			CN27	Youjidami
	CN28	Yalujiang 3 xi			CN28	Wuchangxiangmi
	CN29	Daohuaxiang			CN29	Zhanglixiangmi
	CN30	Youjidami			CN30	Dongbeidami

Figure S1. Spectra of targeted lysoPCs for rice discrimination

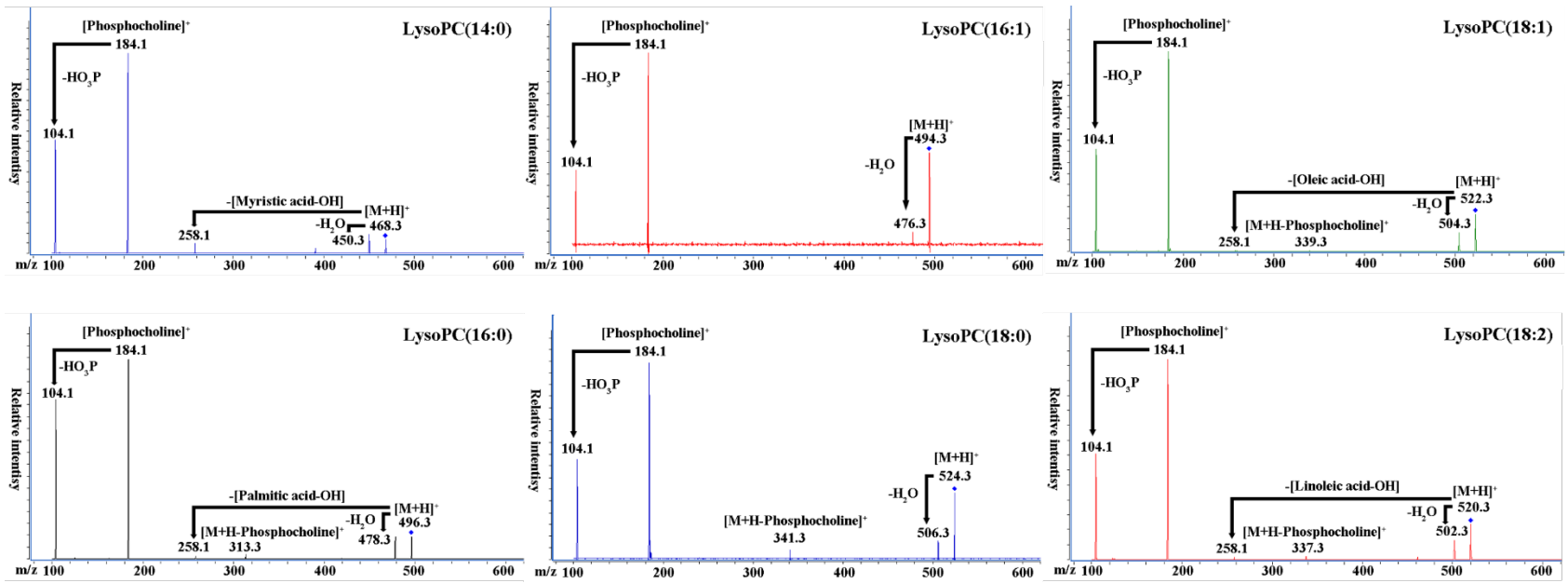


Figure S2. Spectra of targeted lysoPEs for rice discrimination

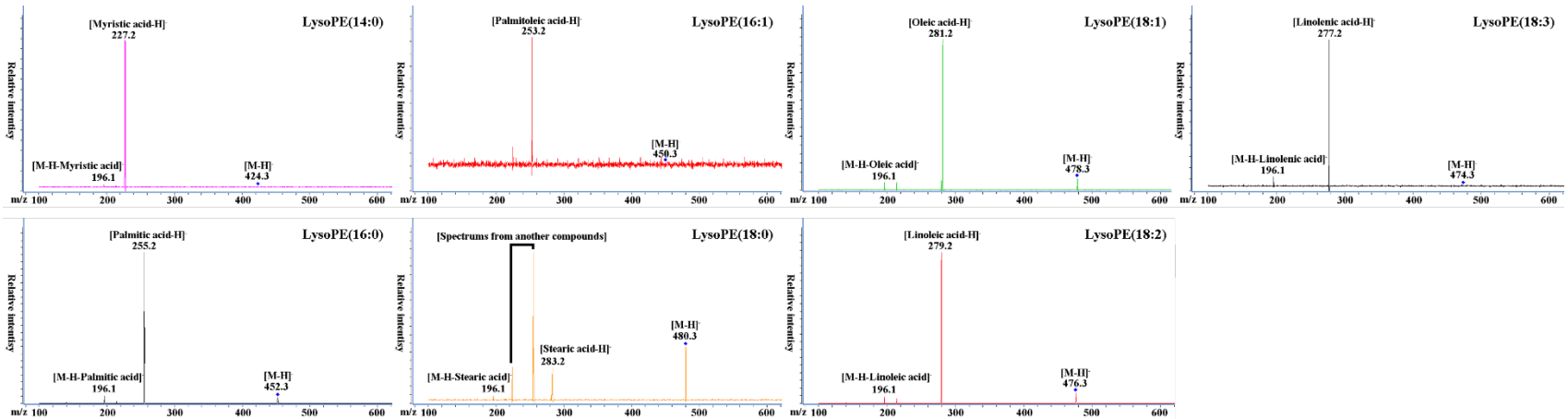
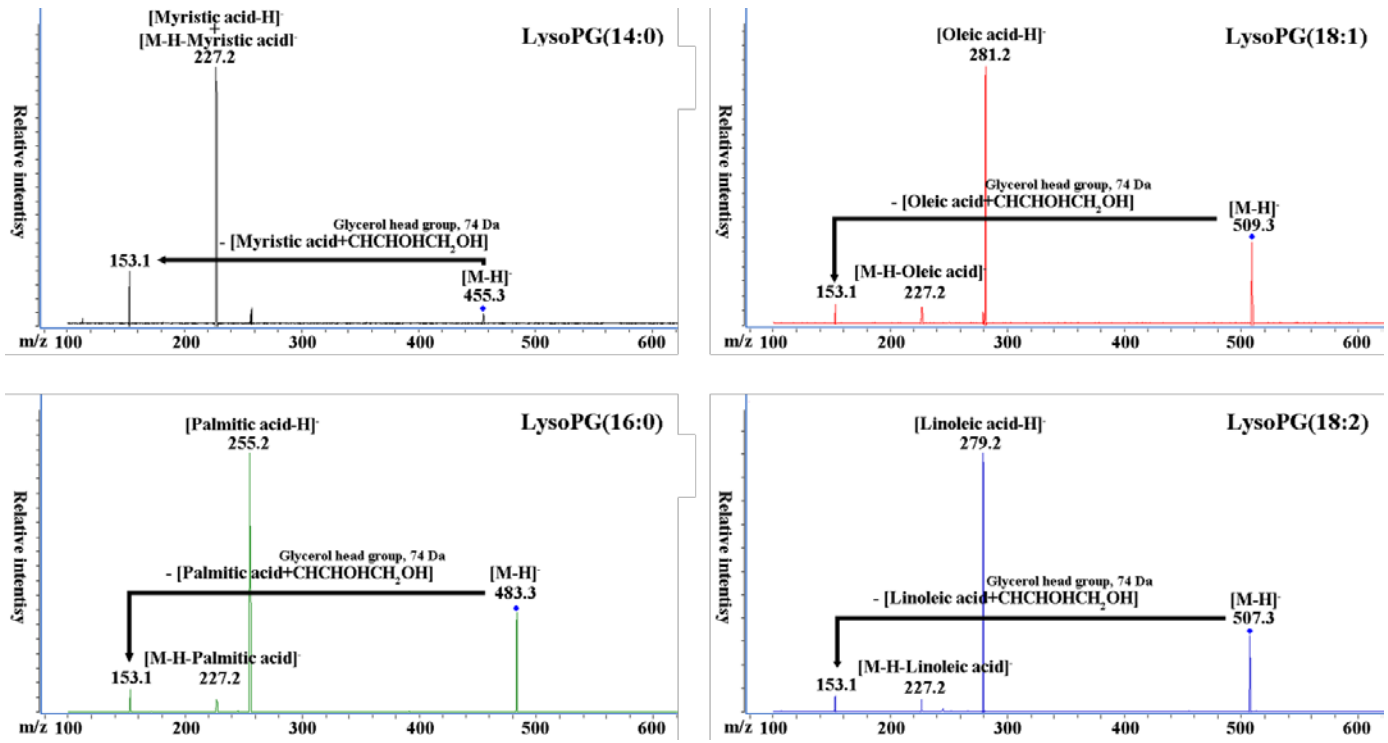
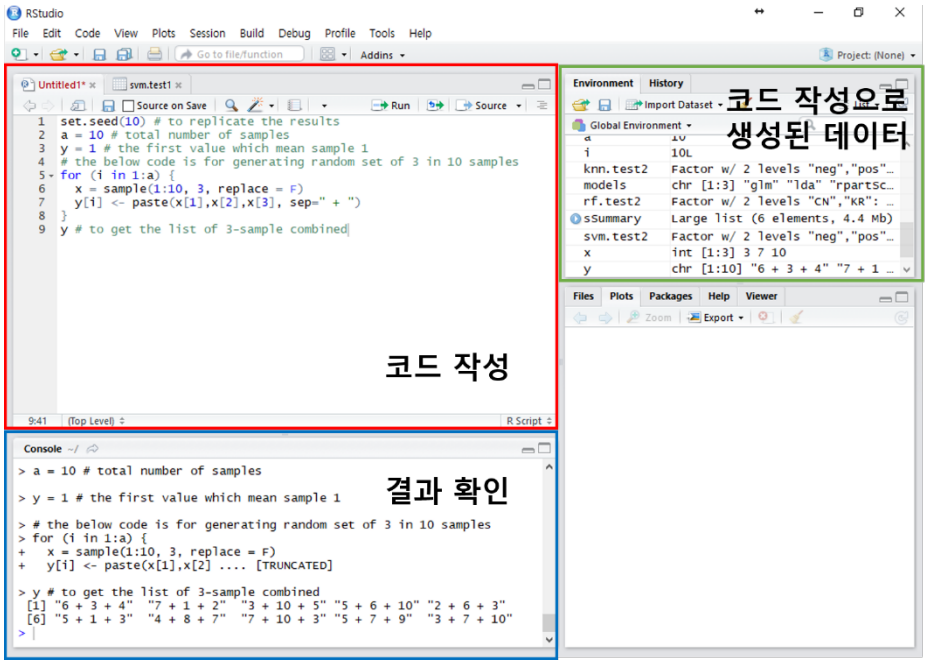


Figure S3. Spectra of targeted lysoPGs for rice discrimination



File S1. R based code for making random manner for blended rice

R 기반의 코드는 RStudio (version 1.0.143)을 이용하여 구동하였다. RStudio 의 기본적인 기능은 아래와 같다.



알고리즘의 구동을 위한 코드를 작성하고 Alt+Cntl+Enter 를 입력하면 코드를 통한 결과값 및 코드 작성으로 인해 생성된 데이터가 형성된다. 혼합 쌀의 제작을 위한 랜덤한 숫자 배열을 생성하기 위한 코드는 아래와 같다.

```
# 시드 설정
      set.seed(10)
```

- 시드를 생성하는 것으로 괄호 안의 숫자를 다르게 입력할 경우 다른 숫자 배열이 생성됨.

```
# 샘플 수 입력
      a = 10

# 샘플의 시작 숫자 입력
```

```
      y = 1

# 샘플의 배열 제작
      for (i in 1:a) {
      x = sample(1:10, 3, replace = F)
      y[i] <- paste(x[1],x[2],x[3], sep=" + ")
      }

# 결과 확인
      y
```

- 위의 코드를 이용하였을 때 최종적으로 아래와 같은 결과값을 얻을 수 있음. 필요에 따라 시드 숫자를 변경하여 사용할 수 있음.

[1] "6 + 3 + 4" [2] "7 + 1 + 2" [3] "3 + 10 + 5" [4] "5 + 6 + 10" [5] "2 + 6 + 3" [6] "5 + 1 + 3" [7] "4 + 8 + 7" [8] "7 + 10 + 3" [9] "5 + 7 + 9" [10] "3 + 7 + 10"

File S2. Data treatment procedures (LC-MS)

모든 LC-MS 분석 결과는 .mzdata 포멧으로 저장한 후 MZmine 2.23.을 이용해 전처리를 진행하였다. 본 연구에서는 아래와 같은 단계 및 패러미터를 이용하여 쌀과 인삼 분석 결과 데이터를 전처리 하였다. 서로 다른 두 종류의 장비를 이용하였기에 재현성, 분석 감도, base line, noise level 등의 수치가 차이가 있으므로 데이터 종류에 따라 다른 패러미터를 적용하여야 한다.

1. Mass detection: 각각의 데이터에서 mass spectrum 을 확인하는 과정으로, base line 의 설정 및 noise 를 제거하는 단계.

Mass detection	쌀	인삼
MS level	1	1
Mass detector	Centroid	Centroid
Noise level	1xE5	2.5xE5

2. Chromatogram builder: mass detection 과정에서 확인한 스팩트럼을 수치로 변환하는 단계.

Chromatogram builder	쌀	인삼
Min time span	0.1 min	0.2 min
Min height	1xE5	2.5xE5
m/z tolerance	50 ppm	200 ppm

3. Chromatogram deconvolution: 한 peak 내에 존재하는 많은 수의 스팩트럼을 각각 하나의 변수로써 적용될 수 있도록 분리하는 단계.

Chromatogram deconvolution	쌀	인삼
Min peak height	5xE3	2xE4
Peak duration range	0.01~2min	0.01~2min
Baseline level	1xE3	1xE4

4. RANSAC aligner: 데이터 간의 머무름 시간, base line 의 높이 등을 통일시켜 모든 데이터를 하나로 합치는 과정.

RANSAC aligner	쌀	인삼
m/z tolerance	50 ppm	200 ppm
RT tolerance	0.05 min	0.1 min
RT tolerance after correlation	0.1 min	0.2 min
RANSAC iterations	100000	100000
Min number of points	20%	20%
Threshold value	0.1 min	0.2 min

5. Gap filling: 데이터를 통합한 후 값이 없는 부분(결측값)을 채워 넣는 단계.

Gap filling	쌀	인삼
m/z tolerance	100 ppm	300 ppm

6. Export to .CSV: 전처리된 데이터를 엑셀 파일 형식으로 만드는 단계.

File S3. R based code for machine learning algorithm.

기계 학습 알고리즘을 통한 판별 모델 제작을 위해서 R 기반의 패키지들을 설치 및 알고리즘을 사용하였다.

```
# 패키지 설치
        install.packages(c("modeval","mlbench","caret"))
```

- Modeval, melbench 및 caret 패키지를 설치.

```
# 패키지 로드
        library(modeval)
        library(mlbench)
        library(caret)
```

- 각각의 패키지를 로딩.

```
# 데이터 로드
        data(Riceclassification)
```

- 엑셀 형식의 데이터를 로드함. 데이터 명은 임의로 정할 수 있으며 본 연구에서는 'Riceclassification' 을 사용.

```
# 데이터 스플리팅 (트레이닝 셋, 테스트 셋)
        set.seed(0704)
```

- 시드를 생성하는 것으로 괄호 안의 숫자를 다르게 입력할 경우 다른 샘플 배열이 생성.

```
        inTraining <-
        createDataPartition(Riceclassification$rice, p = 2/3, list
        = FALSE)
```

- 생성된 시드를 바탕으로 트레이닝 셋, 테스트 셋으로 데이터 파티션을 진행. p=2/3 은 트레이닝 셋의 비율을 67%, 테스트 셋의 비율을 나머지 p=1/3 (33%)으로 한다는 의미.

```
        training <- Riceclassification[inTraining,]
        testing <- Riceclassification[-inTraining,]
        x <- training[,-9]
        y <- training[,9]
        x_test <- testing[,-9]
```

```
y_test <- testing[, 9]
```

- 생성된 트레이닝 셋 및 테스트 셋의 명칭을 간략화 하는 단계.

```
suggest_transformation(x)
```

- Data transformation 방법 중 최선의 방법을 찾는 단계로, 본 연구에서는 Yeo-Johnson 이 선정.

```
# 트레이닝 셋을 이용한 판별 모델 제작

sSummary <- add_model(sSummary, x, y, c("knn",
"nnet", "qda"), modelTag = "Nonlinear", tf="tf2",
tuneLength = 10)
sSummary <- add_model(sSummary, x, y, c("rf",
"rpart", "treebag"), modelTag = "TreeBased", tf="tf2",
tuneLength = 10)
sSummary <- add_model(sSummary, x, y,
c("svmLinear", "svmRadial", "svmPoly"), modelTag =
"svmFamily", tf="tf2", tuneLength = 10)
```

- 다양한 기계 학습 알고리즘을 이용하여 판별 모델 제작. 위의 코드를 이용하여 유사한 방식의 머신 러닝 알고리즘 (Nonlinear=knn, nnet, qda; Tree based= rf, rpart, treebag; svmfamily=svmLinear, svmRadial, svmPoly)을 그룹화 하여 동시에 구동 시킬 수 있으며 모델 제작, 평가 시간을 대폭 감소시킬 수 있음.

```
# 기계 학습 알고리즘 간의 효율 비교 (Visualization)
suggest_auc(sSummary, time = TRUE, "TreeBased")
```

- "" 안에 알고리즘의 그룹 명을 변경하여 알고리즘 간의 효율 비교.
- Receiver operating characteristic (ROC) curve 의 Area under the curve (AUC) 기반으로 각 알고리즘의 퍼포먼스를 평가하고 최적의 모델을 제시함 (AUC 가 1 에 가까울수록 우수한 모델).

```
# 알고리즘 별 cut-off 선정
suggest_probCut(sSummary, "pos")
suggest_probCut(sSummary, "pos", modelTag =
"svmFamily")
```

- 선정된 최적 모델 ROC curve 의 cut-off value 를 통해 해당 모델의 sensitivity, specificity 를 확인함. 이를 통해 기계 학습 알고리즘 간의 퍼포먼스 비교가 가능.

```
# 모델의 gain 및 lift 확인
        suggest_gain(sSummary, "pos", modelTag =
        "svmFamily")
        suggest_gain(sSummary, "pos", type = "Gain") 또는
        suggest_gain(sSummary, "pos", type = "Lift") 또는
        suggest_gain(sSummary, "pos", type = "PctAcc") 또는
        suggest_gain(sSummary, "pos", type = "Pct")
```

- 모델 평가를 위한 패러미터인 gain 및 lift 확인을 위한 다양한 코드를 사용 가능

```
# 판별 모델을 위한 기계 학습 알고리즘 선정 및 테스트 셋 적용을
통한 판별 능력 평가
        knn.test1 <- predict(sSummary$knn_tf2_Nonlinear,
        newdata = x_test, type = "prob")
        knn.test2 <- predict(sSummary$knn_tf2_Nonlinear,
        newdata = x_test)
        confusionMatrix(knn.test2, y_test)

        rf.test1 <- predict(sSummary$rf_tf2_TreeBased,
        newdata = x_test, type = "prob")
        rf.test2 <- predict(sSummary$rf_tf2_TreeBased,
        newdata = x_test)
        confusionMatrix(rf.test2, y_test)
```

- 위의 코드는 k-nearest neighbor 및 random forest 로 생성된 판별 모델에 테스트 셋을 적용하기 위한 것임.
- 테스트 셋 적용 후, accuracy, sensitivity, specificity, positive prediction value, negative prediction value 값을 얻을 수 있음.

File S3. Interpret machine learning results

Random forest 와 Support vector machine 의 두가지 예를 들어 설명, 나머지 기계 학습 알고리즘도 동일한 방식으로 해석할 수 있다.

<100%K vs K/25%C>

Random forest

File S2 의 모델 제작 과정을 진행하면 아래와 같은 데이터가 도출된다.

40 samples
13 predictors
 2 classes: 'A', 'B'

- 40 개의 트레이닝 셋 내의 샘플 중 다시 2/3 을 자체 트레이닝 셋으로 구분하고 1/3 을 자체 테스트 셋으로 설정함. 이후 13 개의 predictors(변수, 여기에서는 lysoPLs)를 이용하여 모델 제작 및 모델 평가(validation) 진행. 해당 과정은 테스트 셋을 적용하기 전 최적의 모델을 제작하기 위함임.
- A 는 한국 쌀, B 는 중국 쌀.

Pre-processing: centered (13), scaled (13)
Resampling: Cross-Validated (5 fold, repeated 3 times)
Summary of sample sizes: 32, 32, 32, 32, 32, 32, ...
Resampling results across tuning parameters:

mtry	ROC	Sens	Spec
1	0.9458333	0.8833333	0.8500000
2	0.9375000	0.9166667	0.8000000
5	0.9583333	0.9666667	0.8333333
6	0.9500000	0.9666667	0.8166667
7	0.9583333	0.9333333	0.8500000
10	0.9666667	0.9333333	0.8666667
11	**0.9708333**	**0.9333333**	**0.8500000**

ROC was used to select the optimal model using the largest value.
The final value used for the model was mtry = 11.

- 트레이닝 셋을 이용한 모델 제작 과정 중, 자동적으로 모델 벨리데이션 과정이 진행됨. 다양한 데이터 normalization, cross-validation 을 통해 임의로 다수의 모델을 제작한 후 sensitivity, specificity 기반의 ROC curve 및 cut-off 값을 도출 하여 최적의 모델을 확보.

Confusion Matrix and Statistics

```
          Reference
Prediction  A B
         A 8 3
         B 2 7

               Accuracy : 0.75
                 95% CI : (0.509, 0.9134)
P-Value [Acc > NIR] : 0.02069

            Sensitivity : 0.8000
            Specificity : 0.7000
         Pos Pred Value : 0.7273
         Neg Pred Value : 0.7778
             Prevalence : 0.5000
         Detection Rate : 0.4000
   Detection Prevalence : 0.5500
      Balanced Accuracy : 0.7500

       'Positive' Class : A
```

- 모델 제작 이후 20 개의 테스트 셋을 적용함. 위와 같이 적용 결과, accuracy, sensitivity, specificity 와 같은 모델 평가를 위한 패러미터 값을 확보할 수 있음.

Support vector machine

마찬가지로 File S2 의 모델 제작 과정을 진행하면 아래와 같은 데이터가 도출된다.

40 samples
13 predictors
 2 classes: 'A', 'B'

Pre-processing: centered (13), scaled (13)
Resampling: Cross-Validated (5 fold, repeated 3 times)
Summary of sample sizes: 32, 32, 32, 32, 32, 32, ...
Resampling results across tuning parameters:

sigma	C	ROC	Sens	Spec
0.01263747	9.453257e+00	0.9125000	0.9333333	0.8166667
0.01603859	7.452380e+00	0.9125000	0.9000000	0.8333333
0.02219815	3.623044e-02	0.1083333	0.1500000	0.1666667
0.02368055	1.997966e-01	0.8916667	0.8500000	0.8333333
0.02802167	6.650491e+00	0.9208333	0.9166667	0.8333333
0.04611226	8.447141e+00	0.9291667	0.9166667	0.8166667
0.07868608	6.147596e-01	0.9083333	0.8500000	0.8666667
0.08953489	7.816633e+01	0.9333333	0.9166667	0.8500000
0.09875071	1.010494e+03	0.9333333	0.9000000	0.8500000
0.12820073	**3.065753e+00**	**0.9541667**	**0.9166667**	**0.9000000**

ROC was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.1282007 and C = 3.065753.

Confusion Matrix and Statistics

	Reference	
Prediction	A	B
A	7	0
B	3	10

```
               Accuracy : 0.85
                 95% CI : (0.6211, 0.9679)
    P-Value [Acc > NIR] : 0.001288

            Sensitivity : 0.7000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.7692
             Prevalence : 0.5000
         Detection Rate : 0.3500
   Detection Prevalence : 0.3500
      Balanced Accuracy : 0.8500

       'Positive' Class : A
```

- Suppor vector machine 알고리즘을 통해 random forest 와 거의 동일한 형식의 데이터를 얻을 수 있음.

  - 단 기계 학습 알고리즘 별로 최적의 ROC 를 구하기 위한 패러미터는 상이하며 (RF=randomly chosen variables (mtry); SVM=kernel smoothing parameter (sigma) 및 cost (C); C5.0=trials, model 및 winnow; NNet=size, decay 및 bag; kNN=the number of closest training examples (k)), 최적의 판별 모델 제작을 위한 패러미터 수치의 높고 낮음을 통해 판별 모델의 우수함을 평가할 수는 없음. 즉 모델 평가를 위해서는 ROC curve, AUC 및 cut-off value 의 sensitivity, specificity 를 이용하는 것이 가장 유용함.

File S4. Definition of accuracy, sensitivity, specificity, positive predictive value, and negative predictive value

	Condition Positive (Korean rice)	Condition Negative (Blended rice)	
Test outcome positive (Detected as Korean rice)	True positive (TP) Korean rice -> Detected as Korean rice	False positive (FP) Blended rice -> Detected as Korean rice	Positive predictive value (PPV) =TP/(TP+FP)
Test outcome Negative (Detected as Blended rice)	False negative (FN) Korean rice -> Detected as Blended rice	True negative (TN) Blended rice -> Detected as Blended rice	Negative predictive value (NPV) = TN/(FN+TN)
	Sensitivity = TP/(TP+FN)	Specificity = TN/(FP+TN)	

Accuracy: Overall correct prediction (TP+TN/TP+TN+FP+FN)

Sensitivity: The proportion of Positive that are accurately identified

Specificity: The proportion of Negative that are accurately identified

PPV: Collection rate of positive test (Korean rice/Detected as Korean rice)

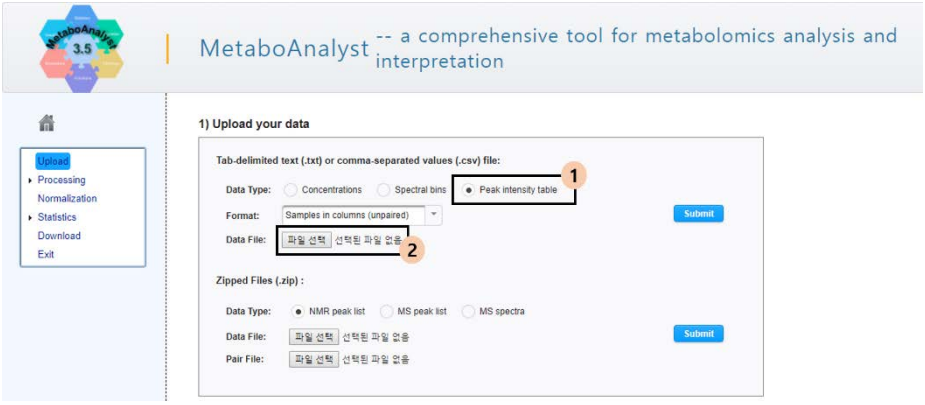
NPV: Collection rate of negative test (Blended rice/Detected as Blended rice)

File S5. Application of multivariate statistical analysis

Principal component analysis (PCA) 및 partial least square-discriminant analysis (PLS-DA)로 대표되는 다변량 분석을 진행하기 위해 metaboanalyst 를 이용한다. 구체적인 순서는 다음과 같다.

1. Metaboanalyst.ca 접속

2. Data type: peak intensity table 선택
   Format: samples in columns (unpaired) 선택



- 사용한 엑셀 파일의 형식

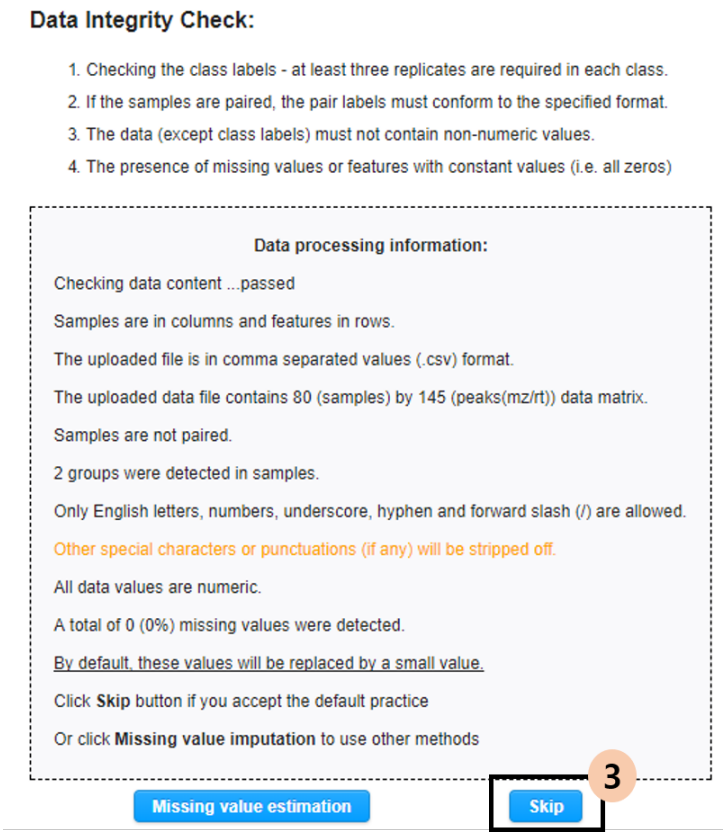
Name of samples

Labeling

Name of variables

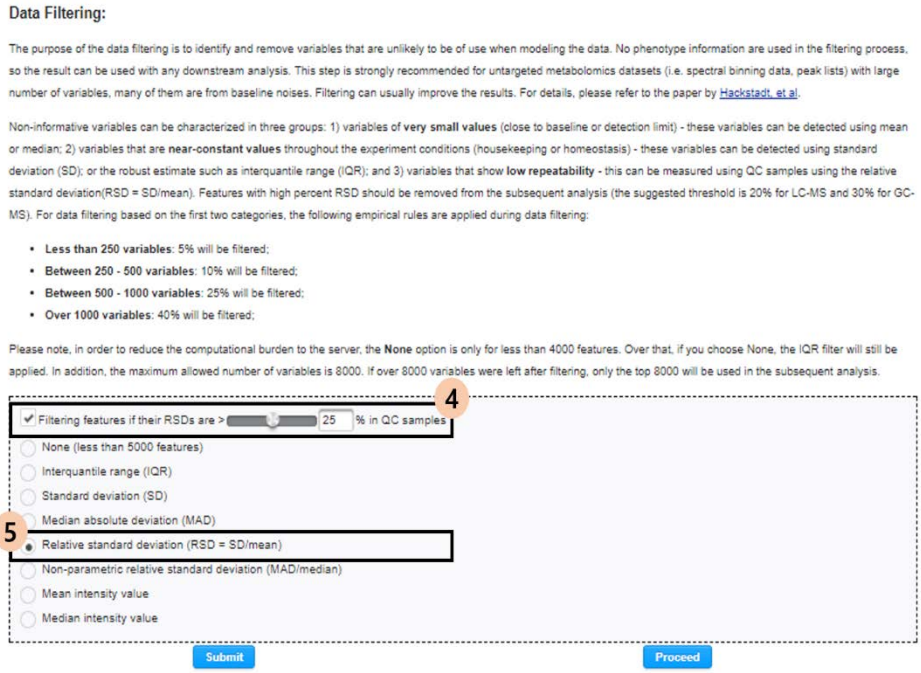
NAME	1A	1B	1C	2A	2B	2C
LABEL	B	C	D	B	C	D
1lysoPC(1	6561	6540	6372	6403	6041	6121
2lysoPC(1	32933	31533	29839	31888	29683	28879
3lysoPC(1	64997	63246	61339	64801	63446	61724
4lysoPC(1	177995	177982	182395	181184	177680	179410
5lysoPC(1	582	596	577	617	640	597
6lysoPC(1	12106	12022	12029	13395	12975	12926
7lysoPG(1	2512	2702	2944	2522	2531	2317
8lysoPE(1	195	225	230	190	226	241
9lysoPE(1	1475	1565	1631	1513	1693	1958

3. Missing value estimation 을 통한 결측치 보정



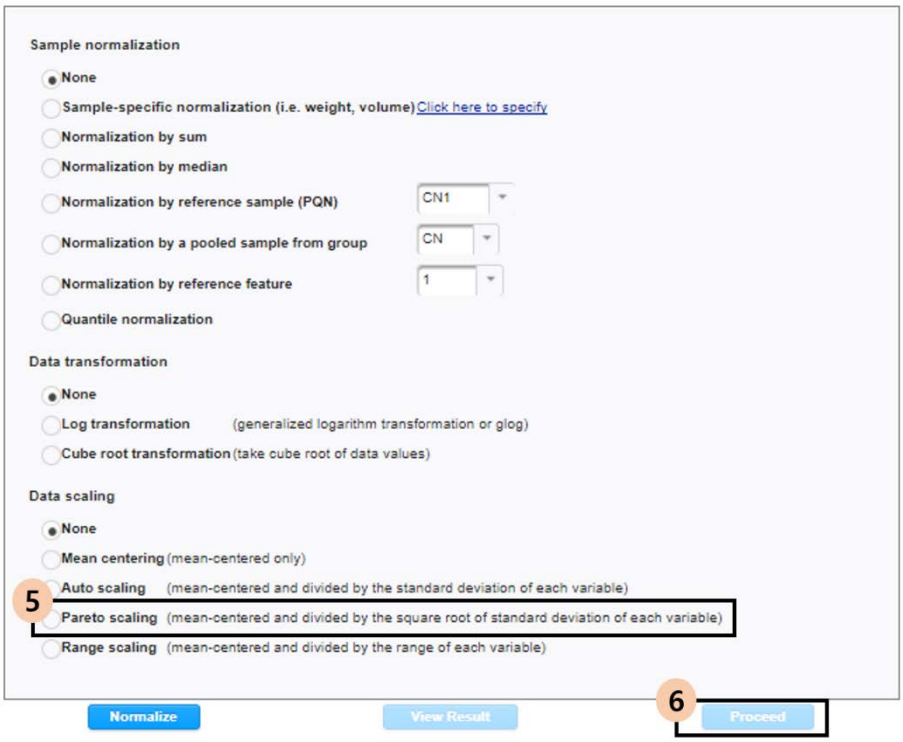
결측치는 항목에 대한 값이 존재하지 않는 것을 말하며, 판별 연구에 다양한 악영향을 끼칠 수 있으므로 데이터 처리 과정에서 제거하거나 작은 수로 대체되어야 함. 본 연구에서는 사전에 결측치의 제거 작업을 진행하였기에 결측치 보정을 따로 진행하지 않음.

4. Data filtering



모델링 제작 시 불필요할 것으로 판단 되는 변수를 제거하기 위한 작업으로 interquantile range (IQR), standard deviation (SD), median absolute deviation (MAD), relative standard deviation (RSD) 등의 기법을 통해서 모델링에 불필요한, 악영향을 끼칠 수 있는 변수들을 제외할 수 있음. 본 연구에서는 QC sample 을 이용하여 25% 이상의 RSD 를 가지는 변수들을 제거함.

5. Normalization

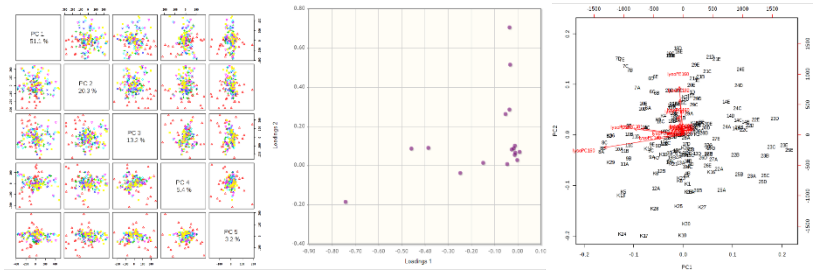


데이터를 보정하는 단계로써 1)내부 표준품을 이용한 데이터 보정, 2) 데이터 transformation, 3) 데이터 scaling 과정을 포함함. 데이터 transformation 은 데이터를 로그 또는 cube root 형태로 변환시키는 것이며, 데이터 scaling 은 데이터를 변환, 전체 자료의 분포를 평균 0, 분산 1 이 되도록 재배치 하는 것임. Transformation 및 scaling 의 목적은 변수의 중요성을 동일하게 평가하기 위한 과정이라 할 수 있음. 본 연구에서는 내부 표준품인 caffeine 을 이용하여 자체적으로 데이터 보정을 진행하였으며, transformation 은 진행하지 않고 pareto scaling (mean-centered and divided by the square root of standard deviation of each variables)을 이용하여 데이터 보정을 진행함.

6. Statistics

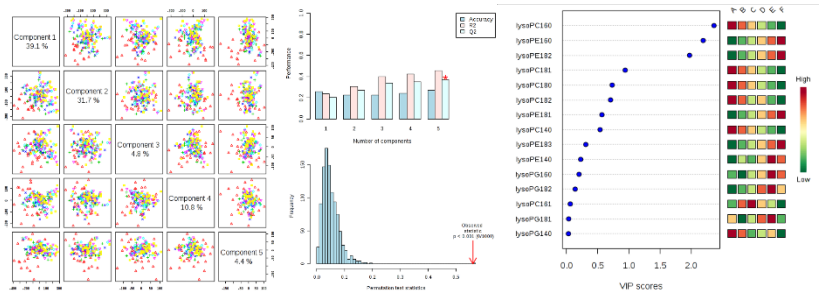
데이터 보정 이후 t-test, ANOVA 와 같은 단변량 분석(univariate analysis)를 비롯하여 PCA, PLS-DA, OPLS-DA 와 같은 다변량 분석(multivariate analysis) 및 random forest 와 같은 기계 학습 기법을 적용하는 것이 가능함.

7. PCA



principal component 에 따른 스코어 플롯들(score plot), 로딩 플롯(loading plot) 및 바이 플롯(biplot)을 확인할 수 있음.

8. PLS-DA



PCA 와 마찬가지로 component 에 따른 스코어 플롯들을 확인할 수 있음. 특히 PLS-DA 에서는 교차검정(cross-validation) 및 순열검정법(permutation test)를 통해 모델을 평가하는 것이 가능함. 이러한 검정을 통해 모델의 accuracy, goodness of fit (R^2), goodness of prediction (Q^2) 및 permutation test 의 p-value 를 얻을 수 있음. 또한 특정 변수가 판별에 미치는 영향을 variable importance in projection (VIP) score 를 통해 측정하는 것이 가능함.

File S6. Phospholipids in rice

백미는 전분, 단백질 및 소량의 지질로 구성되어 있으며, 주로 쌀의 겨(bran, 19.4%~25.5%) 및 쌀 눈(germ, 34.1%~36.5%)에 포함되어 있다. 지질은 특유의 화학 구조에 따라 acylglycerol, free fatty acid, wax, phospholipid (PL), glycolipid 등으로 구분되는데 이러한 지질 중에서 PL 은 유기생물에서 지질 이중층(lipid bilayer)를 형성하거나 에너지원으로서 사용되는 등 중요한 역할을 담당한다. 쌀의 지질 함량 중 대략 10%정도를 차지하는 PL 은 glycerophospholipid (GPL) 및 sphingophospholipid (SPL)의 두가지 카테고리로 구분되는데 쌀에서는 GPL 의 존재만이 확인되었다. 흥미롭게도 GPL 은 전분 중 아밀로오스 성분과 결합을 통해 전분-지질(starch-lipid complex)을 형성하여 질감, 물 흡수 능력, 점착성, gelatinization 온도, 경도, 냄새와 같은 쌀의 품질에 커다란 영향을 미친다 [42]. 특히 이러한 GPL 의 하위 카테고리인 lysophospholipid (lysoPL)은 쌀의 주요 지질로써 아밀로오스와 결합하는 전분-지질의 50% 이상을 차지하고 있으며 1%의 lysoPL 만이 전분과 결합하지 않은 형태로 존재한다. 주로 head group 에 choline 이 결합된 형태인 lysophosphatidylcholine (lysoPC) 및 head group 에 ethanolamine 이 결합된 lysophosphatidylethanolamine (lysoPE)이 백미에 존재하는 주요한 전분-지질의 형태로 알려져 있다. 특히 쌀의 전분에서 발견되는 또 다른 대표적인 전분-지질 성분인 지방산(fatty acid)의 형태는 palmitic acid(C16:0, 48~63%), linoleic acid(C18:2, 25~42%), oleic acid(C18:2, 5%), myristic acid(C14:0, 5%), stearic acid(C18:0, 1%)이며 해당 지방산의 탄소 및 이중결합 숫자가 lysoPL 에서 동일하게 발견되었다 [43, 44]. 즉 본 연구에서 발견된 lysoPLs 는 전분-지질이라 할 수 있다.

전분-지질은 아밀로오스와 결합하여 전분 과립(starch granule)을 형성한다. 아밀로오스는 나선형의 소수성 튜브 형태를 가지고 있으며 지질이 아밀로오스 튜브의 공동 부분에 소수성 fatty acid carbon chain 을 결합한 상태로 존재한다. 따라서 친수성인 지질의 head group 은 튜브 바깥쪽으로 빠져나온 형태이며 이를 아밀로오스-지질 복합체(amylose-lipid complexation)이라 칭한다 [45]. 전분-지질은 아밀로오스가 아밀로펙틴의 형태로 변환되는 것을 방지하는 동시에 아밀로오스가 분해되는 것을 방해하고 결과적으로 쌀의 전분 특성을 결정하는 매우 중요한 성분이다. 한편 전분 생합성을 담당하는 효소를 생성하는 유전자가 고온의 쌀 재배 조건에서 활성이 저해된다는 연구 결과가 발표되었으며 이는 다시 말해 쌀의 등숙 기간(ripening period) 동안의 온도 조건이 쌀의 전분, 특히 아밀로오스의 농도를 결정한다고 할 수 있다. [46, 47]. 이에 따라, non-waxy 타입 쌀의 경우 아밀로오스 함량이 14~18%인 점을 고려하였을 때, 재배 온도의 차이는 아밀로오스 농도 뿐만이 아니라 이와 결합하는 전분-지질의 농도에도 영향을 미칠 것으로 예측할 수 있다. 실제로 Mano, et al 의 연구는 추운 환경에서 재배된 쌀 및 따뜻한 환경에서 재배된 쌀을 대상으로 전분-지질의 또 다른 대표적인 카테고리 중 하나인 불포화 지방산(polyunsaturated fatty acid, PUFA)의 농도를 측정하였고 추운 환경에서 재배된 쌀이 따뜻한 환경에서 재배된 쌀 보다 훨씬 높은 함량의 PUFA 농도를 가짐을 증명하였다 [48]. 또한 식물 또는 조류의 PLs 분석 연구에 따르면 재배 온도 조건이 cholinephosphotrasferase 와 같은 특정 효소의 활성을 변화시킴으로써 저온 조건이 PLs 의 농도를 증가시킴을 확인하였다 [49, 50]. 본 연구의 분석 대상인 한국 쌀 및 중국 쌀의 lysoPLs 를 타겟으로 한 분석 결과에 따르면 lysoPE, lysoPG 는 중국 쌀에서 높은 농도를, lysoPC 는 한국 쌀에서 높은 농도를 보여주고 있다. Head group 종류에 따라 성분의 변화가 달라지는 등 이전 연구에 비해서 좀 더 복잡한 양상을 보이기는 하지만, 쌀의 lysoPLs 의

농도 차이 또한 한국이 중국에 비해 상대적으로 높은 재배 온도를 가지고 있기 때문에 발생한 것으로 추정된다. 향후 재배 온도에 따른 쌀의 lysoPLs 농도 변화를 구체적으로 설명하기 위해서 lysoPLs 을 대상으로 하는 타겟 분석과 함께, 쌀의 lysoPLs 생합성 경로 및 이에 영향을 미치는 효소 등의 연구가 추가적으로 필요할 것이다.

Publications

1. **Lim, Dong Kyu**, et al. "Simultaneous Profiling of Lysoglycerophospholipids in Rice (Oryza sativa L.) Using Direct Infusion-Tandem Mass Spectrometry with Multiple Reaction Monitoring." *Journal of Agricultural and Food Chemistry* 65.12 (2017): 2628-2634. (**First author**)

2. **Lim, Dong Kyu** and Long, Nguyen Phuoc, et al. "Development and assessment of a lysophospholipid-based deep learning model to discriminate geographical origins of white rice." *Scientific Reports* 7.1 (2017): 8552. (**Co-first author**)

3. **Lim, Dong Kyu** and Long, Nguyen Phuoc, et al. "Combination of mass spectrometry-based targeted lipidomics and supervised machine learning algorithms in detecting adulterated admixtures of white rice." *Food Research International* 100 (2017): 814-821. (**Co-first author**)

4. **Lim, Dong Kyu**, et al. "A rapid and reliable method for discriminating rice products from different regions using MCX-based solid-phase extraction and DI-MS/MS-based metabolomics approach." *Journal of Chromatography B* 1061 (2017): 185-192. (**First author**)

5. **Lim, Dong Kyu**, et al. "The integration of multiplatform MS-based metabolomics and multivariate analysis for the geographical origin discrimination of Oryza sativa L." *Journal of Food and Drug Analysis* (2017). (**First author**)

6. **Lim, Dong Kyu**, et al. "Non-destructive profiling of volatile organic compounds using HS-SPME/GC-MS and its application for the geographical discrimination

of white rice." *Journal of Food and Drug Analysis* (2017). (First author)

7. Lim, Dong Kyu, et al. "Optimized mass spectrometry-based metabolite extraction and analysis for the geographical discrimination of white rice (Oryza sativa L.): a method comparison study." *Journal of AOAC International* (2017). (First author)

8. Lim, Dong Kyu, et al. "Impact of Milling on Rice Constituents (Oryza Sativa L.): A Metabolomic Approach." *Analytical Letters* 50.16 (2017): 2519-2529. (First author)